AN OVERVIEW OF PROBABILITY AND STATISTICS

Patrick Boily[1,2,3], Jen Schellinck[2,4,5]

Abstract

Data analysis is sometimes presented in a "point-and-click manner," with tutorials often bypassing foundations in probability and statistics to focus on software use and specific datasets. While modern analysts do not always need to fully understand the theory underpinning the methods that they use, understanding some of the basic concepts can only lead to long-term benefits. In this document, we introduce some of the crucial mathematical notions that will help analysts get the most out of their data.

Keywords

Probability theory, Bayes' theorem, discrete distributions, continuous distributions, joint distributions, random variables, descriptive statistics, sampling distributions, central limit theorem, statistical inference, confidence intervals, hypothesis testing.

Acknowledgement

Parts of the contents and examples were influenced by Rafal Kulik's *Probability and Statistics for Engineers* course notes. Many of the examples are taken from the references.

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada
[2]Data Action Lab, Ottawa, Canada
[3]Idlewyld Analytics and Consulting Services, Wakefield, Canada
[4]Sysabee, Ottawa, Canada
[5]Institute of Cognitive Science, Carleton University, Ottawa, Canada
Email: pboily@uottawa.ca

Contents

In [2], U. Dudley describes **probability theory** as a

> "lovely, coherent whole proceeding from a few axioms with theorems both pretty and deep, and all the more admirable for being applicable."

He then goes on to claim of **statistical theory** that it

> "strikes many mathematicians as being a patchwork of this and that, with ad hock solutions to isolated problems, no unity, and no beauty beyond that of a steel girder bridge: it does its job in a utilitarian fashion, but the rivets show."

In this overview, we introduce the **basic notions** of probability and statistics using a naïve approach (that is to say, without referring to measure theory).

It should not be seen as a formal training replacement for mathematics and statistics students; we present only the bare minimum required for data analysts not to be led astray by machine learning and data science methods. Proofs will be few and far between – the reader interested in more detail is directed to standard references, such as [6,7,10–13,15,16,18,21], from which some of the examples come.

1. Introduction to Probability Theory

Probability theory is the mathematical discipline relating to the numerical description of the likelihood of an event.

1.1 Sample Spaces and Events

Throughout, we will deal with **random experiments** (e.g. measurements of speed/ weight, number and duration of phone calls, etc.).

For any "experiment," the **sample space** is defined as the set of all its **possible outcomes**, often denoted by the symbol $\mathscr{S}$. A sample space can be **discrete** or **continuous**.

An **event** is a collection of outcomes from the sample space $\mathscr{S}$. Events will be denoted by A, B, E_1, E_2, etc.

Examples

- Toss a fair coin – the corresponding (discrete) sample space is $\mathscr{S} = \{\text{Head, Tail}\}$.
- Roll a die – the corresponding (discrete) sample space is $\mathscr{S} = \{1,2,3,4,5,6\}$, with various events represented by
  - rolling an even number: $\{2,4,6\}$;
  - rolling a prime number: $\{2,3,5\}$.
- Suppose we measure the weight (in grams) of a chemical sample – the (continuous) sample space can be represented by $\mathscr{S} = (0,\infty)$, the positive half line, and various events by subsets of $\mathscr{S}$, such as
  - sample is less than 1.5 grams: $(0,1.5)$;
  - sample exceeds 5 grams: $(5,\infty)$.

For any events $A, B \subseteq \mathscr{S}$:

- the **union** $A \cup B$ of A and B are all outcomes in $\mathscr{S}$ contained in either A or B;
- the **intersection** $A \cap B$ of A and B are all outcomes in $\mathscr{S}$ contained in both A and B;
- the **complement** A^c of A (sometimes denoted $\overline{A}$ or $-A$) is the set of all outcomes in $\mathscr{S}$ that are **not** in A.

If A and B have no outcomes in common, they are **mutually exclusive**; which is denoted by $A \cap B = \varnothing$ (the empty set). In particular, A and A^c are always mutually exclusive.[1]

Example

- Roll a die and let $A = \{2,3,5\}$ (a prime number) and $B = \{3,6\}$ (multiples of 3). Then $A \cup B = \{2,3,5,6\}$, $A \cap B = \{3\}$ and $A^c = \{1,4,6\}$.
- 100 plastic samples are analyzed for scratch and shock resistance.

		shock resistance	
		high	low
scratch	high	70	4
resistance	low	1	25

If A is the event that a sample has high shock resistance and B is the event that a sample has high scratch residence, then $A \cap B$ consists of 70 samples.

1.2 Counting Techniques

A **two-stage procedure** can be modeled as having k bags, with m_1 items in the first bag, …, m_k items in k-th bag.

The first stage consists of picking a bag, and the second stage consists of drawing an item out of that bag. This is equivalent to picking one of the $m_1 + \cdots + m_k$ total items.

If all the bags have the same number of items

$$m_1 = \cdots = m_k = n,$$

then there are kn items in total, and this is the total number of ways the two-stage procedure can occur.

Examples

- How many ways are there to first roll a die and then draw a card from a (shuffled) 52–card pack?

  Answer: there are 6 ways the first step can turn out, and for each of these (the stages are independent, in fact) there are 52 ways to draw the card. Thus there are $6 \times 52 = 312$ ways this can turn out.

- How many ways are there to draw two tickets numbered 1 to 100 from a bag, the first with the right hand and the second with the left hand?

  Answer: There are 100 ways to pick the first number; for *each of these* there are 99 ways to pick the second number. Thus $100 \times 99 = 9900$ ways.

[1] Events can be represented graphically using Venn diagrams – mutually exclusive events are those which do not have a common intersection.

Multi-Stage Procedures A k-**stage process** is a process for which:

- there are n_1 possibilities at stage 1;
- regardless of the 1st outcome there are n_2 possibilities at stage 2,
- ...
- regardless of the previous outcomes, there are n_k choices at stage k.

There are then

$$n_1 \times n_2 \cdots \times n_k$$

total ways the process can turn out.

1.3 Ordered Samples

Suppose we have a bag of n billiard balls numbered $1, \ldots, n$. We can draw an **ordered sample** of size r by picking balls from the bag:

- **with replacement**, or
- **without replacement**.

With how many different collection of r balls can we end up in each of those cases (each is an r-stage procedure)?

Key Notion: all the object (balls) can be differentiated (using numbers, colours, etc.)

Sampling With Replacement (Order Important) If we replace each ball into the bag after it is picked, then every draw is the same (there are n ways it can turn out).

According to our earlier result, there are

$$\underbrace{n \times n \times \cdots \times n}_{r \text{ stages}} = n^r$$

ways to select an ordered sample of size r **with replacement** from a set with n objects $\{1, 2, \ldots, n\}$.

Sampling Without Replacement (Order Important) If we **do not** replace each ball into the bag after it is drawn, then the choices for the second draw depend on the result of the first draw, and there are only $n-1$ possible outcomes.

Whatever the first two draws were, there are $n-2$ ways to draw the third ball, and so on.

Thus there are

$$\underbrace{n \times (n-1) \times \cdots \times (n-r+1)}_{r \text{ stages}} = {}_nP_r \quad \text{(common symbol)}$$

ways to select an ordered sample of size $r \leq n$ **without replacement** from a set of n objects $\{1, 2, \ldots, n\}$.

Factorial Notation For a positive integer n, write

$$n! = n(n-1)(n-2)\cdots 1.$$

There are two possibilities:

- when $r = n$, ${}_nP_r = n!$, and the ordered selection (without replacement) is called a **permutation**;
- when $r < n$, we can write

$$\begin{aligned} {}_nP_r &= \frac{n(n-1)\cdots(n-r+1)}{} \frac{(n-r)\cdots 1}{(n-r)\cdots 1} \\ &= \frac{n!}{(n-r)!} = n \times \cdots \times (n-r+1). \end{aligned}$$

By convention, we set $0! = 1$, so that

$${}_nP_r = \frac{n!}{(n-r)!}, \quad \text{for all } r \leq n.$$

Examples

- In how many different ways can 6 balls be drawn *in order* without replacement from a bag of balls numbered 1 to 49?

  Answer: We compute

  $${}_{49}P_6 = 49 \times 48 \times 47 \times 46 \times 45 \times 44 = 10,068,347,520.$$

  This is the number of ways the actual drawing of the balls can occur for Lotto 6/49 in real-time (balls drawn one by one).

- How many 6-digits PIN codes can you create from the set of digits $\{0, 1, \ldots, 9\}$?

  Answer: If the digits may be repeated, we see that

  $$10 \times 10 \times 10 \times 10 \times 10 \times 10 = 10^6 = 1,000,000.$$

  If the digits may not be repeated, we have instead

  $${}_{10}P_6 = 10 \times 9 \times 8 \times 7 \times 6 \times 5 = 151,200.$$

1.4 Unordered Samples

Suppose now that we **cannot** distinguish between different ordered samples; when we look up the Lotto 6/49 results in the newspaper, for instance, we have no way of knowing the order in which the balls were drawn:

$$1-2-3-4-5-6$$

could mean that the first drawn ball was ball # 1, the second drawn ball was ball # 2, etc., but it could also mean that the first ball drawn was ball # 4, the second one, ball # 3, etc., or any other combinations of the first 6 balls.

Denote the (as yet unknown) number of unordered samples of size r from a set of size n by ${}_nC_r$. We can derive the expression for ${}_nC_r$ by noting that the following two processes are equivalent:

- take an **ordered** sample of size r (there are ${}_nP_r$ ways to do this);
- take an **unordered** sample of size r (there are ${}_nC_r$ ways to do this) **and then** rearrange (permute) the objects in the sample (there are $r!$ ways to do this).

Thus

$$ {}_nP_r = {}_nC_r \times r! \quad \Longrightarrow \quad {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{(n-r)!\,r!} = \binom{n}{r}. $$

This last notation is called a **binomial coefficient** (read as "n-choose-r") and is commonly used in textbooks.

Example: in how many ways can the "Lotto 6/49 draw" be reported in the newspaper (where they are always reported in increasing order)?

Answer: this number is the same as the number of *unordered samples* of size 6 (different re-orderings of same 6 numbers are indistinguishable), so

$$ \begin{aligned} {}_{49}C_6 &= \binom{49}{6} = \frac{49\times 48\times 47\times 46\times 45\times 44}{6\times 5\times 4\times 3\times 2\times 1} \\ &= \frac{10,068,347,520}{720} = 13,983,816\,. \end{aligned} $$

There exists a variety of binomial coefficient identities, such as

$$ \begin{aligned} \binom{n}{k} &= \binom{n}{n-k}, \quad \text{for all } 0 \le k \le n, \\ \sum_{k=0}^{n} \binom{n}{k} &= 2^n, \quad \text{for all } 0 \le n, \\ \binom{n+1}{k+1} &= \binom{n}{k} + \binom{n}{k+1}, \quad \text{for all } 0 \le k \le n-1 \\ \sum_{j=k}^{n} \binom{j}{k} &= \binom{n+1}{k+1}, \quad \text{for all } 0 \le n, \text{ etc..} \end{aligned} $$

1.5 Probability of an Event

For situations where we have a random experiment which has exactly N possible **mutually exclusive**, **equally likely** outcomes, we can assign a probability to an event A by counting the number of outcomes that correspond to A – its **relative frequency**.

If that count is a, then

$$ P(A) = \frac{a}{N}. $$

The probability of each individual outcome is thus $1/N$.

Examples

- Toss a fair coin – the sample space is $\mathscr{S} = \{\text{Head, Tail}\}$, i.e. $N = 2$. The probability of observing a Head on a toss is thus $\frac{1}{2}$.
- Throw a fair six sided die. There are $N = 6$ possible outcomes. The sample space is

  $$ \mathscr{S} = \{1,2,3,4,5,6\}. $$

  If A corresponds to observing a multiple of 3, then $A = \{3,6\}$ and $a = 2$, so that

  $$ \text{Prob(number is a multiple of 3)} = P(A) = \frac{2}{6} = \frac{1}{3}. $$

- The probabilities of seeing an even/odd number are:

  $$ \begin{aligned} \text{Prob\{even\}} &= P(\{2,4,6\}) = \frac{3}{6} = \frac{1}{2}; \\ \text{Prob\{prime\}} &= P(\{2,3,5\}) = 1 - P(\{1,4,6\}) = \frac{1}{2}. \end{aligned} $$

- In a group of 1000 people it is known that 545 have high blood pressure. 1 person is selected randomly. What is the probability that this person has high blood pressure?

  Answer: the relative frequency of people with high blood pressure is 0.545.

This approach to probability is called the **frequentist interpretation**. It is based on the idea that the theoretical probability of an event is given by the behaviour of the empirical (observed) relative frequency of the event over long-run repeatable and independent experiments (i.e. when $N \to \infty$).

This is the classical definition, and the one used in this document, but there are competing interpretations which may be more appropriate depending on the context; chiefly, the **Bayesian interpretation** (see [3, 9] for details) and the **propensity interpretation** (introducing causality as a mechanism).

Axioms of Probability The modern definition of probability is **axiomatic** (according to Kolmogorov's seminal work).

The **probability of an event** $A \subseteq \mathscr{S}$ is a numerical value satisfying the following properties:

1. for any event A, $1 \ge P(A) \ge 0$;
2. for the complete sample space $\mathscr{S}$, $P(\mathscr{S}) = 1$;
3. for the empty event $\varnothing$, $P(\varnothing) = 0$, and
4. for two **mutually exclusive** events A and B, the probability that A or B occurs is $P(A \cup B) = P(A) + P(B)$.

Since $\mathscr{S} = A \cup A^c$, and A and A^c are mutually exclusive, then

$$ \begin{aligned} 1 &\overset{\text{A2}}{=} P(\mathscr{S}) = P(A \cup A^c) \overset{\text{A4}}{=} P(A) + P(A^c) \\ &\Longrightarrow P(A^c) = 1 - P(A). \end{aligned} $$

Examples

- Throw a single six sided die and record the number that is shown. Let A and B be the events that the number is a multiple of or smaller than 3, respectively. Then $A = \{3,6\}$, $B = \{1,2\}$ and A and B are mutually exclusive since $A \cap B = \varnothing$. Then

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) = \frac{2}{6} + \frac{2}{6} = \frac{2}{3}.$$

- An urn contains 4 white balls, 3 red balls and 1 black ball. Draw one ball, and denote the following events by $W = \{\text{the ball is white}\}$, $R = \{\text{the ball is red}\}$ and $B = \{\text{the ball is black}\}$. Then

$$P(W) = 1/2, \quad P(R) = 3/8, \quad P(B) = 1/8,$$

and $P(W \text{ or } R) = 7/8$.

General Addition Rule This useful rule is a direct consequence of the axioms of probability:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Example: an electronic gadget consists of two components, A and B. We know from experience that $P(A \text{ fails}) = 0.2$, $P(B \text{ fails}) = 0.3$ and $P(\text{both } A \text{ and } B \text{ fail}) = 0.15$. Find $P(\text{at least one of } A \text{ and } B \text{ fails})$ and $P(\text{neither } A \text{ nor } B \text{ fails})$.

Answer: write A for "A fails" and similarly for B. Then we are looking to compute

$$\begin{aligned} P(\text{at least one fails}) &= P(A \cup B) \\ &= P(A) + P(B) - P(A \cap B) = 0.35\,; \\ P(\text{neither fail}) &= 1 - P(\text{at least one fails}) = 0.65\,. \end{aligned}$$

If A, B are mutually exclusive, $P(A \cap B) = P(\varnothing) = 0$ and

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B).$$

With three events, the addition rule expands as follows:

$$\begin{aligned} P(A \cup B \cup C) = &P(A) + P(B) + P(C) \\ &- P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &+ P(A \cap B \cap C). \end{aligned}$$

1.6 Conditional Probability and Independent Events

Any two events A and B satisfying

$$P(A \cap B) = P(A) \times P(B)$$

are said to be **independent**.[2] When events are not independent, we say that they are **dependent** or **conditional**.

Mutual exclusivity and independence are unrelated concepts. The only way for events A and B to be mutually exclusive **and** independent is for either A or B (or both) to be a non-event (the empty event):

$$\varnothing = P(A \cap B) = P(A) \times P(B) \implies P(A) = 0 \text{ or } P(B) = 0$$
$$\implies A = \varnothing \text{ or } B = \varnothing.$$

Examples

- Flip a **fair** coin twice – the 4 possible outcomes are all equally likely: $\mathscr{S} = \{HH, HT, TH, TT\}$. Let

$$A = \{HH\} \cup \{HT\}$$

denote "head on first flip", $B = \{HH\} \cup \{TH\}$ "head on second flip". Note that $A \cup B \neq \mathscr{S}$ and $A \cap B = \{HH\}$. By the general addition rule,

$$\begin{aligned} P(A) &= P(\{HH\}) + P(\{HT\}) - P(\{HH\} \cap \{HT\}) \\ &= \frac{1}{4} + \frac{1}{4} - P(\varnothing) = \frac{1}{2} - 0 = \frac{1}{2}\,. \end{aligned}$$

Similarly, $P(B) = P(\{HH\}) + P(\{TH\}) = \frac{1}{2}$, and so $P(A)P(B) = \frac{1}{4}$. But $P(A \cap B) = P(\{HH\})$ is also $\frac{1}{4}$, so A and B are independent.

- A card is drawn from a regular well-shuffled 52-card North American deck. Let A be the event that it is an ace and D be the event that it is a diamond. These two events are independent. Indeed, there are 4 aces

$$P(A) = \frac{4}{52} = \frac{1}{13}$$

and 13 diamonds

$$P(D) = \frac{13}{52} = \frac{1}{4}$$

in such a deck, so that

$$P(A)P(D) = \frac{1}{13} \times \frac{1}{4} = \frac{1}{52}\,,$$

and exactly 1 ace of diamonds in the deck, so that $P(A \cap D)$ is also $\frac{1}{52}$.

- A six-sided die numbered $1-6$ is loaded in such a way that the probability of rolling each value is *proportional* to that value. Find $P(3)$.

Answer: Let $\mathscr{S} = \{1,2,3,4,5,6\}$ be the value showing after a single toss; for some proportional constant v, we have $P(k) = kv$, for $k \in \mathscr{S}$. By Axiom **A2**, $P(\mathscr{S}) = P(1) + \cdots + P(6) = 1$, so that

$$1 = \sum_{k=1}^{6} P(k) = \sum_{k=1}^{6} kv = v \sum_{k=1}^{6} k = v\frac{(6+1)(6)}{2} = 21v\,.$$

Hence $v = 1/21$ and $P(3) = 3v = 3/21 = 1/7$.

[2] This is a purely mathematical definition, but it agrees with the intuitive notion of independence in simple examples.

- Now the die is rolled twice, the second toss *independent* of the first. Find $P(3_1, 3_2)$.

  Answer: the experiment is such that $P(3_1) = 1/7$ and $P(3_2) = 1/7$, as seen in the previous example. Since the die tosses are independent, then

  $$P(3_1 \cap 3_2) = P(3_1)P(3_2) = 1/49\,.$$ [3]

- Is a 2-engine plane more likely to be forced down than a 3-engine plane?

  Answer: this question is easier to answer if we assume that **engines fail independently** (this is no doubt convenient, but the jury is still out as to whether it is realistic). In what follows, let p be the probability that an engine fails.[4]

  The next step is to decide what type engine failure will force a plane down:[5]

  - A 2-engine plane will be forced down if both engines fail – the probability is p^2;
  - A 3-engine plane will be forced down if any pair of engines fail, or if all 3 fail.
    - **Pair:** the probability that exactly 1 pair of engines will fail independently (i.e. two engines fail and one does not) is

      $$p \times p \times (1-p).$$

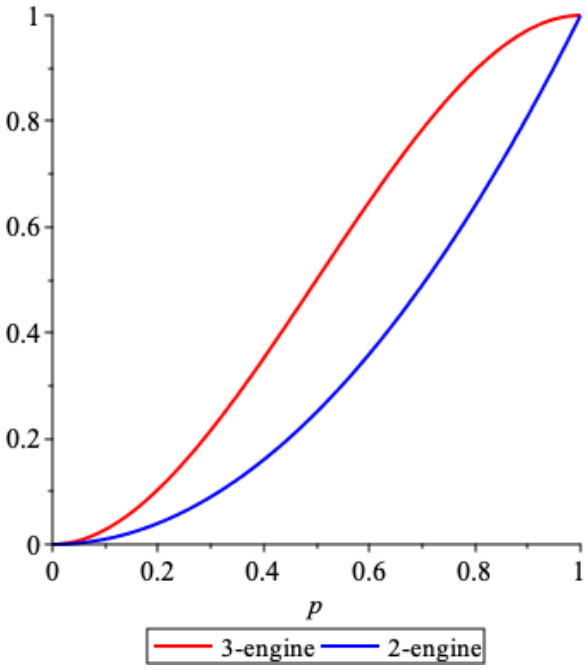
      The order in which the engines fail does not matter: there are ${}_3C_2 = \frac{3!}{2!1!} = 3$ ways in which a pair of engines can fail: for 3 engines A, B, C, these are AB, AC, BC.
    - **All 3:** the probability of all three engines failing independently is p^3.

  The probability ≥ 2 engines failing is thus

  $$P(2+ \text{ engines fail}) = 3p^2(1-p)+p^3 = 3p^2-2p^3.$$

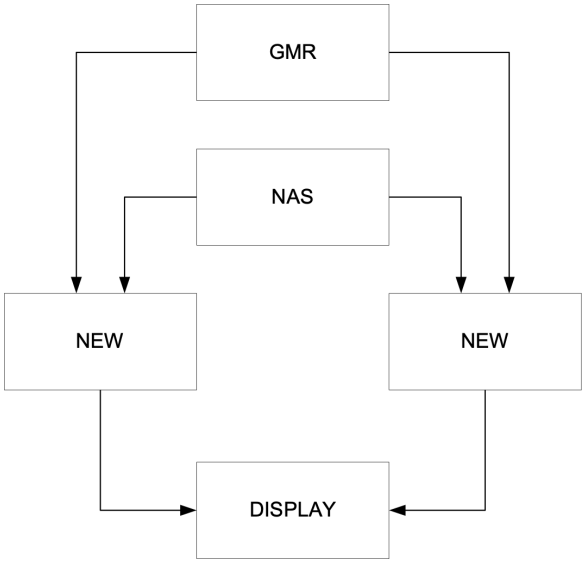
  Basically it's safer to use a 2-engine plane than a 3-engine plane: the 3-engine plane will be forced down more often, assuming it needs 2 engines to fly.

  This "makes sense": the 2-engine plane need 50% of its engines working, while the 3-engine plane needs 66% (see the image at the top of the column on the right to get a sense of what the probabilities are for $0 \leq p \leq 1$).



- (Taken from [14]) Air traffic control is a safety-related activity – each piece of equipment is designed to the highest safety standards and in many cases duplicate equipment is provided so that if one item fails another takes over.

  A new system is to be provided passing information from Heathrow Airport to Terminal Control at West Drayton. As part of the system design a decision has to be made as to whether it is necessary to provide duplication. The new system takes data from the *Ground Movements Radar* (GMR) at Heathrow, combines this with data from the *National Airspace System NAS*, and sends the output to a display at *Terminal Control*.



[3] Is it clear what is meant by "independent tosses"?
[4] What are some realistic values of p?
[5] There is nothing to that effect in the problem statement, so we have to make another set of assumptions.

For all existing systems, records of failure are kept and an experimental probability of failure is calculated annually using the previous 4 years.

The **reliability** of a system is defined as $R = 1 - P$, where $P = P(\text{failure})$.

Given: $R_{\text{GMR}} = R_{\text{NAS}} = 0.9999$ (i.e. 1 failure in $10,000$ hours).

Assumption: the components' failure probabilities are independent.

For the system above, if a single NEW module is introduced the reliability of the system (STD – **single thread design**) is

$$R_{\text{STD}} = R_{\text{GMR}} \times R_{\text{NEW}} \times R_{\text{NAS}}.$$

If the NEW module is duplicated, the reliability of this **dual thread design** is

$$R_{\text{DTD}} = R_{\text{GMR}} \times (1-(1-R_{\text{NEW}})^2) \times R_{\text{NAS}}.$$

Duplicating the NEW module causes an improvement in reliability of

$$\rho = \frac{R_{\text{DTD}}}{R_{\text{STD}}} = \frac{(1-(1-R_{\text{NEW}})^2)}{R_{\text{NEW}}} \times 100\%.$$

For the NEW module, no historical data is available. Instead, we work out the improvement achieved by using the dual thread design for various values of R_{NEW}.

R_{NEW}	0.1	0.2	0.5	0.75
ρ (%)	190	180	150	125
R_{NEW}	0.99	0.999	0.9999	0.99999
ρ (%)	101	100.1	100.01	100.001

If the NEW module is very unreliable (i.e. R_{NEW} is small), then there is a significant benefit in using the dual thread design (ρ is large).[6]

If the new module is as reliable as NAS and GMR, that is, if

$$R_{\text{GMR}} = R_{\text{NEW}} = R_{\text{NAS}} = 0.9999,$$

then the single thread design has a combined reliability of 0.9997 (i.e. 3 failures in $10,000$ hours), whereas the dual thread design has a combined reliability of 0.9998 (i.e. 2 failures in $10,000$ hours).

If the probability of failure is independent for each component, we could conclude from this that the reliability gain from a dual thread design probably does not justify the extra cost.

In the last two examples, we had to make **additional assumptions** in order to answer the questions.

[6] But why would we install a module which we know to be unreliable in the first place?

Conditional Probability It is easier to understand independence of events through the **conditional probability** of an event B given that another event A has occurred, defined by as

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}.$$

Note that this definition only makes sense when "A can happen" i.e. $P(A) > 0$. If $P(A)P(B) > 0$, then

$$P(A \cap B) = P(A) \times P(B \mid A) = P(B) \times P(A \mid B) = P(B \cap A);$$

A and B are thus independent if $P(B \mid A) = P(B)$ and $P(A \mid B) = P(A)$.

Examples

- From a group of 100 people, 1 is selected. What is the probability that this person has high blood pressure (HBP)?

  Answer: if we know nothing else about the population, this is an **(unconditional) probability**, namely

  $$P(\text{HBP}) = \frac{\#\text{individuals with HBP in the population}}{100}.$$

- If instead we first filter out all people with low cholesterol level, and then select 1 person. What is the probability that this person has HBP?

  Answer: this is the **conditional probability**

  $$P(\text{HBP} \mid \text{high cholesterol});$$

  the probability of selecting a person with HBP, given high cholesterol levels, presumably different from $P(\text{HBP} \mid \text{low cholesterol})$.

- A sample of 249 individuals is taken and each person is classified by blood type and tuberculosis (TB) status.

	O	**A**	**B**	**AB**	Total
TB	34	37	31	11	113
no TB	55	50	24	7	136
Total	89	87	55	18	249

  The (unconditional) probability that a random individual has TB is $P(\text{TB}) = \frac{\#\text{TB}}{249} = \frac{113}{249} = 0.454$. Among those individuals with type **B** blood, the (conditional) probability of having TB is

  $$P(\text{TB} \mid \text{type } \mathbf{B}) = \frac{P(\text{TB} \cap \text{type } \mathbf{B})}{P(\text{type } \mathbf{B})} = \frac{31}{55} = \frac{31/249}{55/249} = 0.564.$$

- A family has two children (not twins). What is the probability that the youngest child is a girl given that at least one of the children is a girl? Assume that boys and girls are equally likely to be born.

Answer: let A and B be the events that the youngest child is a girl and that at least one child is a girl, respectively:

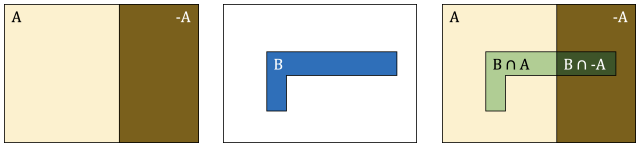
$$A = \{\text{GG}, \text{BG}\} \quad \text{and} \quad B = \{\text{GG}, \text{BG}, \text{GB}\},$$

so that $A \cap B = A$. Then $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{2/4}{3/4} = \frac{2}{3}$ (and not $\frac{1}{2}$, as might naively be believed).

Incidentally, $P(A \cap B) = P(A) \neq P(A) \times P(B)$, which means that A and B are **not** independent events.

Law of Total Probability Let A and B be two events. From set theory, we have

$$B = (A \cap B) \cup (\bar{A} \cap B).$$



Note that $A \cap B$ and $\bar{A} \cap B$ are mutually exclusive, so that, according to Axiom **A4**, we have

$$P(B) = P(A \cap B) + P(\bar{A} \cap B).$$

Now, assuming that $\varnothing \neq A \neq \mathscr{S}$,

$$P(A \cap B) = P(B \mid A)P(A) \quad \text{and} \quad P(\bar{A} \cap B) = P(B \mid \bar{A})P(\bar{A}),$$

so that

$$P(B) = P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A}).$$

This generalizes as follows: if $A_1, \ldots A_k$ are **mutually exclusive** and **exhaustive** (i.e. $A_i \cap A_j = \varnothing$ for all $i \neq j$ and $A_1 \cup \ldots \cup A_k = \mathscr{S}$), then for any event B

$$\begin{aligned} P(B) &= \sum_{j=1}^{k} P(B \mid A_j)P(A_j) \\ &= P(B \mid A_1)P(A_1) + \ldots + P(B \mid A_k)P(A_k). \end{aligned}$$

Example: use the **Law of Total Probability** (rule above) to compute $P(\text{TB})$ using the data from the previous example.

Answer: the blood types $\{\mathbf{O}, \mathbf{A}, \mathbf{B}, \mathbf{AB}\}$ form a mutually exclusive partition of the population, with

$$P(\mathbf{O}) = \frac{89}{249}, \; P(\mathbf{A}) = \frac{87}{249}, \; P(\mathbf{B}) = \frac{55}{249}, \; P(\mathbf{AB}) = \frac{18}{249}.$$

It is easy to see that $P(\mathbf{O}) + P(\mathbf{A}) + P(\mathbf{B}) + P(\mathbf{AB}) = 1$. Furthermore,

$$\begin{aligned} P(\text{TB} \mid \mathbf{O}) &= \tfrac{P(\text{TB} \cap \mathbf{O})}{P(\mathbf{O})} = \tfrac{34}{89}, \; P(\text{TB} \mid \mathbf{A}) = \tfrac{P(\text{TB} \cap \mathbf{A})}{P(\mathbf{A})} = \tfrac{37}{87}, \\ P(\text{TB} \mid \mathbf{B}) &= \tfrac{P(\text{TB} \cap \mathbf{B})}{P(\mathbf{B})} = \tfrac{31}{55}, \; P(\text{TB} \mid \mathbf{AB}) = \tfrac{P(\text{TB} \cap \mathbf{AB})}{P(\mathbf{AB})} = \tfrac{11}{18}. \end{aligned}$$

According to the law of total probability,

$$\begin{aligned} P(\text{TB}) = {} & P(\text{TB} \mid \mathbf{O})P(\mathbf{O}) + P(\text{TB} \mid \mathbf{A})P(\mathbf{A}) \\ & + P(\text{TB} \mid \mathbf{B})P(\mathbf{B}) + P(\text{TB} \mid \mathbf{AB})P(\mathbf{AB}), \end{aligned}$$

so that

$$\begin{aligned} P(\text{TB}) &= \frac{34}{89} \cdot \frac{89}{249} + \frac{37}{87} \cdot \frac{87}{249} + \frac{31}{55} \cdot \frac{55}{249} + \frac{11}{18} \cdot \frac{18}{249} \\ &= \frac{34 + 37 + 31 + 11}{249} = \frac{113}{249} = 0.454, \end{aligned}$$

which matches with the result of the previous example.

1.7 Bayes' Theorem

After an experiment generates an outcome, we are often interested in the probability that a certain condition was present given an outcome (or that a particular hypothesis was valid, say).

We have noted before that if $P(A)P(B) > 0$, then

$$P(A \cap B) = P(A) \times P(B \mid A) = P(B) \times P(A \mid B) = P(B \cap A);$$

this can be re-written as **Bayes' Theorem**:

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}.$$

Bayes' Theorem is a powerful tool in probability analysis, but it is a simple corollary of the rules of probability.

Central Data Analysis Question Given everything that was known prior to the experiment, does the collected/observed data support (or invalidate) the hypothesis/presence of a certain condition?

The **problem** is that this is usually impossible to compute directly. Bayes' Theorem offers a **possible solution**:

$$\begin{aligned} P(\text{hypothesis} \mid \text{data}) &= \frac{P(\text{data} \mid \text{hypothesis}) \times P(\text{hypothesis})}{P(\text{data})} \\ &\propto P(\text{data} \mid \text{hypothesis}) \times P(\text{hypothesis}), \end{aligned}$$

in which the terms on the right might be easier to compute than the term on the left.

Bayesian Vernacular In Bayes' Theorem:

- $P(\text{hypothesis})$ is the probability of the hypothesis being true prior to the experiment (called the **prior**);
- $P(\text{hypothesis} \mid \text{data})$ is the probability of the hypothesis being true once the experimental data is taken into account (called the **posterior**);
- $P(\text{data} \mid \text{hypothesis})$ is the probability of the experimental data being observed assuming that the hypothesis is true (called the **likelihood**).

The theorem is often presented as posterior $\propto$ likelihood $\times$ prior, which is to say, **beliefs should be updated in the presence of new information**.

Formulations If A, B are events for which $P(A)P(B) > 0$, then Bayes' Theorem can be re-written, using the law of total probability, as

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})},$$

or, in the general case where $A_1, \ldots A_k$ are **mutually exclusive** and **exhaustive** events, then for any event B and for each $1 \leq i \leq k$,

$$\begin{aligned} P(A_i \mid B) &= \frac{P(B \mid A_i)P(A_i)}{P(B)} \\ &= \frac{P(B \mid A_i)P(A_i)}{P(B \mid A_1)P(A_1) + \ldots + P(B \mid A_k)P(A_k)}. \end{aligned}$$

Examples

- In 1999, Nissan sold three car models in North America: Sentra (S), Maxima (M), and Pathfinder (PA). Of the vehicles sold that year, 50% were S, 30% were M and 20% were PA. In the same year 12% of the S, 15% of the M, and 25% of the PA had a particular defect D.

  1. If you own a 1999 Nissan, what is the probability that it has the defect?

     Answer: in the language of conditional probability,

     $$\begin{aligned} P(\mathrm{S}) = 0.5,\ P(\mathrm{M}) = 0.3,\ P(\mathrm{Pa}) = 0.2, \\ P(D \mid \mathrm{S}) = 0.12,\ P(D \mid \mathrm{M}) = 0.15,\ P(D \mid \mathrm{PA}) = 0.25, \end{aligned}$$

     so that

     $$\begin{aligned} P(D) &= P(D \mid \mathrm{S}) \times P(\mathrm{S}) + P(D \mid \mathrm{M}) \times P(\mathrm{M}) \\ &\quad + P(D \mid \mathrm{Pa}) \times P(\mathrm{Pa}) \\ &= 0.12 \cdot 0.5 + 0.15 \cdot 0.3 + 0.25 \cdot 0.2 \\ &= 0.155 = 15.5\%. \end{aligned}$$

  2. If a 1999 Nissan has defect D, what model is it likely to be?

     Answer: in the first part we computed the total probability $P(D)$; in this part, we compare the posterior probabilities $P(\mathrm{M} \mid D)$, $P(\mathrm{S} \mid D)$, and $P(\mathrm{Pa} \mid D)$ (and not the priors!), computed using Bayes' Theorem:

     $$\begin{aligned} P(\mathrm{S} \mid D) &= \tfrac{P(D|\mathrm{S})P(\mathrm{S})}{P(D)} = \tfrac{0.12 \times 0.5}{0.155} \approx 38.7\% \\ P(\mathrm{M} \mid D) &= \tfrac{P(D|\mathrm{M})P(\mathrm{M})}{P(D)} = \tfrac{0.15 \times 0.3}{0.155} \approx 29.0\% \\ P(\mathrm{Pa} \mid D) &= \tfrac{P(D|\mathrm{Pa})P(\mathrm{Pa})}{P(D)} = \tfrac{0.25 \times 0.2}{0.155} \approx 32.3\% \end{aligned}$$

     Even though Sentras are the least likely to have the defect D, their overall prevalence in the population carry them over the hump.

- Suppose that a test for a particular disease has a very high success rate. If a patient
  - has the disease, the test reports a 'positive' with probability 0.99;
  - does not have the disease, the test reports a 'negative' with prob 0.95.

  Assume that only 0.1% of the population has the disease. What is the probability that a patient who tests positive does not have the disease?

  Answer: Let D be the event that the patient has the disease, and A be the event that the test is positive. The probability of a true positive is

  $$\begin{aligned} P(D \mid A) &= \frac{P(A \mid D)P(D)}{P(A \mid D)P(D) + P(A \mid D^c)P(D^c)} \\ &= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.05 \times 0.999} \approx 0.019. \end{aligned}$$

  The probability of a false positive is thus $1 - 0.019 \approx 0.981$. Despite the apparent high accuracy of the test, the incidence of the disease is so low (1 in a 1000) that the vast majority of patients who test positive (98 in 100) do not have the disease.

  The 2 in 100 who are true positives still represent 20 times the proportion of positives found in the population (before the outcome of the test is known).[7]

- (**Monty Hall Problem**) On a game show, you are given the choice of three doors. Behind one of the doors is a prize; behind the others, dirty and smelly rubbish bins.

  You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, behind which is a bin. She then says to you, "Do you want to switch from door No. 1 to No. 2?"

  Is it to your advantage to do so?

[7] It is important to remember that when dealing with probabilities, **both** the likelihood and the prevalence have to be taken into account.

Answer: in what follows, let S and D be the events that switching to another door is a successful strategy and that the prize is behind the original door, respectively.

- Let's first assume that the host opens no door. What is the probability that switching to another door in this scenario would prove to be a successful strategy?

  If the prize is behind the original door, switching would succeed 0% of the time:

  $$P(\mathrm{S} \mid \mathrm{D}) = 0.$$

  Note that the prior is $P(\mathrm{D}) = 1/3$.

  If the prize is not behind the original door, switching would succeed 50% of the time:

  $$P(\mathrm{S} \mid \mathrm{D}^c) = 1/2.$$

  Note that the prior is $P(\mathrm{D}^c) = 2/3$. Thus,

  $$\begin{aligned} P(\mathrm{S}) &= P(\mathrm{S} \mid \mathrm{D})P(\mathrm{D}) + P(\mathrm{S} \mid \mathrm{D}^c)P(\mathrm{D}^c) \\ &= 0 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{2}{3} \approx 33\%. \end{aligned}$$

- Now let's assume that the host opens one of the other two doors to show a rubbish bin. What is the probability that switching to another door in this scenario would prove to be a successful strategy?

  If the prize is behind the original door, switching would succeed 0% of the time:

  $$P(\mathrm{S} \mid \mathrm{D}) = 0.$$

  Note that the prior is $P(\mathrm{D}) = 1/3$.

  If the prize is not behind the original door, switching would succeed 100% of the time:

  $$P(\mathrm{S} \mid \mathrm{D}^c) = 1.$$

  Note that the prior is $P(\mathrm{D}^c) = 2/3$. Thus,

  $$\begin{aligned} P(\mathrm{S}) &= P(\mathrm{S} \mid \mathrm{D})P(\mathrm{D}) + P(\mathrm{S} \mid \mathrm{D}^c)P(\mathrm{D}^c) \\ &= 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} \approx 67\%. \end{aligned}$$

  If no door is opened, switching is not a winning strategy, resulting in success only 33% of the time. If a door is opened, however, switching becomes the winning strategy, resulting in success 67% of the time.

This problem has attracted a lot of attention over the years due to its counter-intuitive result. There is no paradox when one understands conditional probabilities.

2. Discrete Distributions

The principles of probability theory introduced in the previous section are simple, and they are always valid. In this section and the next, we will see how some of the computations can be made easier with the use of distributions.

2.1 Random Variables and Distributions

Recall that, for any random "experiment," the set of all possible outcomes is denoted by $\mathscr{S}$. A **random variable** (r.v.) is a function $X : \mathscr{S} \to \mathbb{R}$, which is to say, it is a rule that associates a (real) number to every outcome of the experiment; $\mathscr{S}$ is the **domain** of the r.v. X and $X(\mathscr{S}) \subseteq \mathbb{R}$ is its **range**.

A **probability distribution function** (p.d.f.) is a function $f : \mathbb{R} \to \mathbb{R}$ which specifies the probabilities of the values in the range $X(\mathscr{S})$.

When $\mathscr{S}$ is **discrete**,[8] we say that X is a **discrete r.v.** and the p.d.f. is called a **probability mass function** (p.m.f.).

Notation Throughout, we use the following notation:

- capital roman letters (X, Y, etc.) denote r.v., and
- corresponding lower case roman letters (x, y, etc.) denote *generic values taken* by the r.v.

A discrete r.v. can be used to **define events**: if X takes values $X(\mathscr{S}) = \{x_i\}$, then we can define events

$$A_i = \{s \in \mathscr{S} : X(s) = x_i\} :$$

- the p.m.f. of X is

  $$f(x) = P\left(\{s \in \mathscr{S} : X(s) = x\}\right) := P(X = x);$$

- its **cumulative distribution function** (c.d.f.) is

  $$F(x) = P(X \leq x).$$

Properties If X is a discrete random variable with p.m.f. $f(x)$ and c.d.f. $F(x)$, then

- $0 < f(x) \leq 1$ for all $x \in X(\mathscr{S})$;
- $\sum_{s \in \mathscr{S}} f(X(s)) = \sum_{x \in X(\mathscr{S})} f(x) = 1$;
- for any event $A \subseteq \mathscr{S}$, $P(X \in A) = \sum_{x \in A} f(x)$;
- for any $a, b \in \mathbb{R}$,

  $$\begin{aligned} P(a < X) &= 1 - P(X \leq a) = 1 - F(a) \\ P(X < b) &= P(X \leq b) - P(X = b) = F(b) - f(b) \end{aligned}$$

- for any $a, b \in \mathbb{R}$,

  $$\begin{aligned} P(a \leq X) &= 1 - P(X < a) \\ &= 1 - (P(X \leq a) - P(X = a)) \\ &= 1 - F(a) + f(a) \end{aligned}$$

[8] For the purpose of this document, a discrete set is one in which all points are **isolated**: $\mathbb{N}$ and finite sets are discrete, but $\mathbb{Q}$ and $\mathbb{R}$ are not.

We can use these results to compute the probability of a **discrete** r.v. X falling in various intervals:

$$\begin{aligned}
P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\
&= F(b) - F(a) \\
P(a \leq X \leq b) &= P(a < X \leq b) + P(X = a) \\
&= F(b) - F(a) + f(a) \\
P(a < X < b) &= P(a < X \leq b) - P(X = b) \\
&= F(b) - F(a) - f(b) \\
P(a \leq X < b) &= P(a \leq X \leq b) - P(X = b) \\
&= F(b) - F(a) + f(a) - f(b)
\end{aligned}$$

Examples

- Flip a fair coin – the outcome space is $\mathscr{S} = \{\text{Head}, \text{Tail}\}$. Let $X : S \to \mathbb{R}$ be defined by $X(\text{Head}) = 1$ and $X(\text{Tail}) = 0$. Then X is a discrete random variable (as a convenience, we write $X = 1$ and $X = 0$).

  If the coin is fair, the p.m.f. of X is $f : \mathbb{R} \to \mathbb{R}$, where

  $$f(0) = P(X = 0) = 1/2,\ f(1) = P(X = 1) = 1/2,$$
  $$f(x) = 0 \text{ for all other } x.$$

- Roll a fair die – the outcome space is $\mathscr{S} = \{1, \ldots, 6\}$. Let $X : \mathscr{S} \to \mathbb{R}$ be defined by $X(i) = i$ for $i = 1, \ldots, 6$. Then X is a discrete r.v.

  If the die is fair, the p.m.f. of X is $f : \mathbb{R} \to \mathbb{R}$, where

  $$f(i) = P(X = i) = 1/6, \text{ for } i = 1, \ldots, 6,$$
  $$f(x) = 0 \text{ for all other } x.$$

- For the random variable X from the previous example, the c.d.f. is $F : \mathbb{R} \to \mathbb{R}$, where

  $$F(x) = P(X \leq x) = \begin{cases} 0 & \text{if } x < 1 \\ i/6 & \text{if } i \leq x < i+1,\ i = 1, \ldots, 6 \\ 1 & \text{if } x \geq 6 \end{cases}$$

- For the same random variable, we can compute the probability $P(3 \leq X \leq 5)$ directly:

  $$\begin{aligned} P(3 \leq X \leq 5) &= P(X = 3) + P(X = 4) + P(X = 5) \\ &= \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2}, \end{aligned}$$

  or we can use the c.d.f.:

  $$P(3 \leq X \leq 5) = F(5) - F(3) + f(3) = \tfrac{5}{6} - \tfrac{3}{6} + \tfrac{1}{6} = \tfrac{1}{2}.$$

- The number of calls received over a specific time period, X, is a discrete random variable, with potential values $0, 1, 2, \ldots$.

- Consider a 5–card poker hand consisting of cards selected at random from a 52–card deck. Find the probability distribution of X, where X indicates the number of red cards (♢ and ♡) in the hand.

  Answer: in all there are $\binom{52}{5}$ ways to select a 5–card poker hand from a 52–card deck. By construction, X can take on values $x = 0, 1, 2, 3, 4, 5$.

  If $X = 0$, then none of the 5 cards in the hands are ♢ or ♡, and all of the 5 cards in the hands are ♠ or ♣. There are thus $\binom{26}{0} \cdot \binom{26}{5}$ 5–card hands that only contain black cards, and

  $$P(X = 0) = \frac{\binom{26}{0} \cdot \binom{26}{5}}{\binom{52}{5}}.$$

  In general, if $X = x$, $x = 0, 1, 2, 3, 4, 5$, there are $\binom{26}{x}$ ways of having x ♢ or ♡ in the hand, and $\binom{26}{5-x}$ ways of having $5 - x$ ♠ and ♣ in the hand, so that

  $$f(x) = P(X = x) = \begin{cases} \frac{\binom{26}{x} \cdot \binom{26}{5-x}}{\binom{52}{5}}, & x = 0, 1, 2, 3, 4, 5; \\ 0 \text{ otherwise} \end{cases}$$

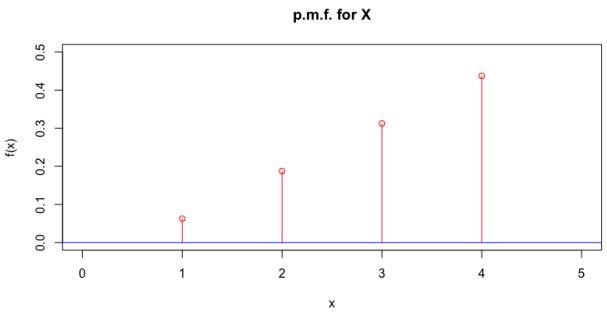
- Find the c.d.f. of a discrete random variable X with p.m.f. $f(x) = 0.1x$ if $x = 1, 2, 3, 4$ and $f(x) = 0$ otherwise.

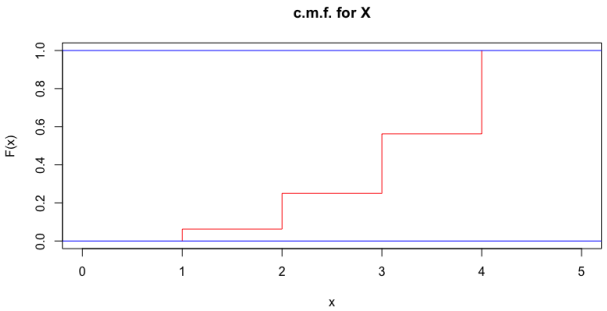
  Answer: $f(x)$ is indeed a p.m.f. as $0 < f(x) \leq 1$ for all x and

  $$\sum_{x=1}^{4} 0.1x = 0.1(1 + 2 + 3 + 4) = 0.1\frac{4(5)}{2} = 1.$$

  Computing $F(x) = P(X \leq x)$ yields

  $$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 0.1 & \text{if } 1 \leq x < 2 \\ 0.3 & \text{if } 2 \leq x < 3 \\ 0.6 & \text{if } 3 \leq x < 4 \\ 1 & \text{if } x \geq 4 \end{cases}$$





2.2 Expectation of a Discrete Random Variable

The **expectation** of a discrete random variable X is

$$\mathrm{E}[X] = \sum_x x \cdot P(X = x) = \sum_x x f(x),$$

where the sum extends over all values of x taken by X.

The definition can be extended to a general function of X:

$$\mathrm{E}[u(X)] = \sum_x u(x) P(X = x) = \sum_x u(x) f(x).$$

As an important example, note that

$$\mathrm{E}[X^2] = \sum_x x^2 P(X = x) = \sum_x x^2 f(x).$$

Examples

- What is the expectation on the roll Z of 6—sided die?

  Answer: if the die is fair, then

$$\begin{aligned}\mathrm{E}[Z] &= \sum_{z=1}^{6} z \cdot P(Z = z) = \frac{1}{6}\sum_{z=1}^{6} z \\ &= \frac{1}{6} \cdot \frac{6(7)}{2} = 3.5.\end{aligned}$$

- For each 1\$ bet in a gambling game, a player can win 3\$ with probability $\frac{1}{3}$ and lose 1\$ with probability $\frac{2}{3}$. Let X be the net gain/loss from the game. Find the expected value of the game.

  Answer: X can take on the value 2\$ for a win and -2\$ for a loss (outcome $-$ bet). The expected value of X is thus

$$\mathrm{E}[X] = 2 \cdot \frac{1}{3} + (-2) \cdot \frac{2}{3} = -\frac{2}{3}.$$

- If Z is the number showing on a roll of a fair 6—sided die, find $\mathrm{E}[Z^2]$ and $\mathrm{E}[(Z-3.5)^2]$.

  Answer:

$$\begin{aligned}\mathrm{E}[Z^2] &= \sum_z z^2 P(Z = z) = \frac{1}{6}\sum_{z=1}^{6} z^2 \\ &= \frac{1}{6}(1^2 + \cdots + 6^2) = \frac{91}{6} \\ \mathrm{E}[(Z-3.5)^2] &= \sum_{z=1}^{6} (z-3.5)^2 P(Z = z) \\ &= \frac{1}{6}\sum_{z=1}^{6} (z-3.5)^2 \\ &= \frac{(1-3.5)^2 + \cdots + (6-3.5)^2}{6} = \frac{35}{12}.\end{aligned}$$

The expectation of a random variable is the average value that it takes.

Mean and Variance We can interpret the expectation as the average or the **mean** of X, which we often denote by $\mu = \mu_X$. For instance, in the example of the fair die,

$$\mu_Z = \mathrm{E}[Z] = 3.5$$

Note that in the final example, we could have written

$$\mathrm{E}[(Z-3.5)^2] = \mathrm{E}[(Z - \mathrm{E}[Z])^2].$$

This is an important quantity associated to a random variable X, its **variance** $\mathrm{Var}[X]$.

The variance of a discrete random variable X is the **expected squared difference from the mean**:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}[(X-\mu_X)^2] = \sum_x (x-\mu_X)^2 P(X = x) \\ &= \sum_x \left(x^2 - 2x\mu_X + \mu_X^2\right) f(x) \\ &= \sum_x x^2 f(x) - 2\mu_X \sum_x x f(x) + \mu_X^2 \sum_x f(x) \\ &= \mathrm{E}[X^2] - 2\mu_X \mu_X + \mu_X^2 \cdot 1 \\ &= \mathrm{E}[X^2] - \mu_X^2.\end{aligned}$$

This is also sometimes written as $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}^2[X]$.

Standard Deviation The **standard deviation** of a discrete random variable X is defined directly from the variance:

$$\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}.$$

The mean is a measure of **centrality** and it gives an idea as to where the **bulk** of a distribution is located; the variance and standard deviation provide information about the **spread** – distributions with higher variance/SD are **more spread out about the average**.

Example: let X and Y be random variables with the following p.d.f.

x	$P(X=x)$
-2	1/5
-1	1/5
0	1/5
1	1/5
2	1/5

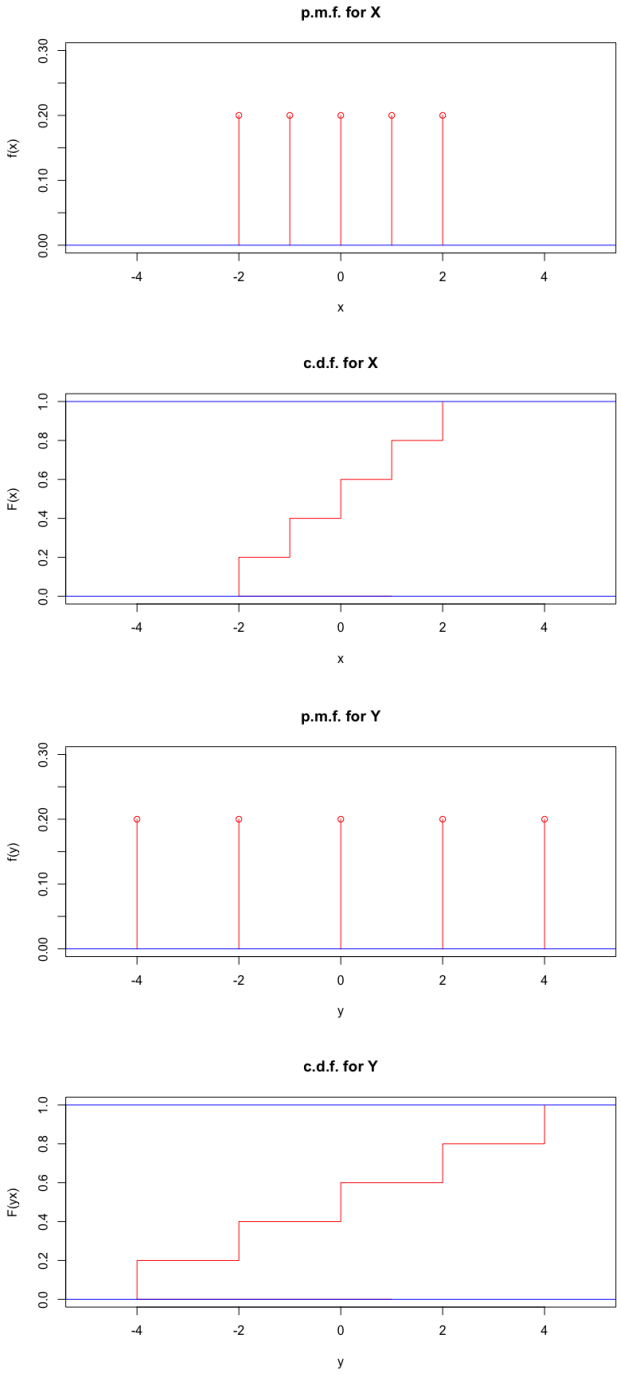
y	$P(Y=y)$
-4	1/5
-2	1/5
0	1/5
2	1/5
4	1/5

Compute the expected values and compare the variances.

Answer: We have $\mathrm{E}[X] = \mathrm{E}[Y] = 0$ and

$$2 = \mathrm{Var}[X] < \mathrm{Var}[Y] = 8,$$

meaning that we would expect both distributions to be centered at 0, but Y should be more spread-out than X.



Properties Let X, Y be random variables and $a \in \mathbb{R}$. Then

- $\mathrm{E}[aX] = a\mathrm{E}[X]$;
- $\mathrm{E}[X+a] = \mathrm{E}[X] + a$;
- $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$;
- in general, $\mathrm{E}[XY] \neq \mathrm{E}[X]\mathrm{E}[Y]$;
- $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$, $\mathrm{SD}[aX] = |a|\,\mathrm{SD}[X]$;
- $\mathrm{Var}[X+a] = \mathrm{Var}[X]$, $\mathrm{SD}[X+a] = \mathrm{SD}[X]$.

2.3 Binomial Distributions

Recall that the number of unordered samples of size r from a set of size n is

$$ {}_nC_r = \binom{n}{r} = \frac{n!}{(n-r)!r!}. $$

Examples

- $2! \times 4! = (1\times 2)\times(1\times 2\times 3\times 4) = 48$, but $(2\times 4)! = 8! = 40320$.
- $\binom{5}{1} = \frac{5!}{1!\times 4!} = \frac{1\times 2\times 3\times 4\times 5}{1\times(1\times 2\times 3\times 4)} = \frac{5}{1} = 5$.
- In general: $\binom{n}{1} = n$ and $\binom{n}{0} = 1$.
- $\binom{6}{2} = \frac{6!}{2!\times 4!} = \frac{4!\times 5\times 6}{2!\times 4!} = \frac{5\times 6}{2} = 15$.
- $\binom{27}{22} = \frac{27!}{22!\times 5!} = \frac{22!\times 23\times 24\times 25\times 26\times 27}{5!\times 22!} = \frac{23\times 24\times 25\times 26\times 27}{120}$.

Binomial Experiments A **Bernoulli trial** is a random experiment with two possible outcomes, “success" and “failure". Let p denote the probability of a success.

A **binomial experiment** consists of n repeated *independent* Bernoulli trials, each with the same probability of success, p.

Examples

- female/male births;
- satisfactory/defective items on a production line;
- sampling with replacement with two types of item,
- etc.

Probability Mass Function In a binomial experiment of n independent events, each with probability of success p, the number of successes X is a discrete random variable that follows a **binomial distribution** with parameters (n, p):

$$ f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \text{for } x = 0, 1, 2, \ldots, n. $$

This is often abbreviated to “$X \sim \mathscr{B}(n, p)$”.

If $X \sim \mathscr{B}(1, p)$, then $P(X = 0) = 1-p$ and $P(X = 1) = p$, so

$$ \mathrm{E}[X] = (1-p)\cdot 0 + p \cdot 1 = p\,. $$

Expectation and Variance If $X \sim \mathscr{B}(n, p)$, it can be shown that

$$ \mathrm{E}[X] = \sum_{x=0}^{n} xP(X = x) = np, $$

and

$$ \mathrm{Var}[X] = \mathrm{E}\left[(X-np)^2\right] = \sum_{x=0}^{n} (x-np)^2 P(X = x) = np(1-p)\,.^{9} $$

Recognizing that certain situations can be modeled *via* a distribution whose p.m.f. and c.d.f. are already known can simplify eventual computations.

[9] We will see an easier way to derive these by interpreting X as a sum of other discrete random variables.

Examples

- Suppose that water samples taken in some well-defined region have a 10% probability of being polluted. If 12 samples are selected independently, then it is reasonable to model the number X of polluted samples as $\mathscr{B}(12, 0.1)$.

  Find

  a) $\mathrm{E}[X]$ and $\mathrm{Var}[X]$;

  b) $P(X = 3)$;

  c) $P(X \leq 3)$.

  Solution:

  a) If $X \sim \mathscr{B}(n, p)$, then

  $$\mathrm{E}[X] = np \quad \text{and} \quad \mathrm{Var}[X] = np(1-p).$$

  With $n = 12$ and $p = 0.1$, we obtain

  $$\mathrm{E}[X] = 12 \times 0.1 = 1.2;$$
  $$\mathrm{Var}[X] = 12 \times 0.1 \times 0.9 = 1.08\,.$$

  b) By definition,

  $$P(X = 3) = \binom{12}{3}(0.1)^3(0.9)^9 \approx 0.0852.$$

  c) By definition,

  $$P(X \leq 3) = \sum_{x=0}^{3} P(X = x) = \sum_{x=0}^{3} \binom{12}{x}(0.1)^x(0.9)^{12-x}.$$

  This sum can be computed directly, however, for $X \sim \mathscr{B}(12, 0.1)$, $P(X \leq 3)$ can also be read directly from tabulated values (see below):

12	0	0.2824	0.0687	0.0138	0.0022	0.0002	0.0000			
	1	0.6590	0.2749	0.0850	0.0196	0.0032	0.0003	0.0000		
	2	0.8891	0.5583	0.2528	0.0834	0.0193	0.0028	0.0002		
	3	0.9744	0.7946	0.4925	0.2253	0.0730	0.0153	0.0017	0.0000	
	4	0.9957	0.9274	0.7237	0.4382	0.1938	0.0573	0.0095	0.0006	
	5	0.9995	0.9806	0.8822	0.6652	0.3872	0.1582	0.0386	0.0039	0.0000
	6	0.9999	0.9961	0.9614	0.8418	0.6128	0.3348	0.1178	0.0194	0.0005
	7	1.0000	0.9994	0.9905	0.9427	0.8062	0.5618	0.2763	0.0726	0.0043
	8		0.9999	0.9983	0.9847	0.9270	0.7747	0.5075	0.2054	0.0256
	9		1.0000	0.9998	0.9972	0.9807	0.9166	0.7472	0.4417	0.1109
	10			1.0000	0.9997	0.9968	0.9804	0.9150	0.7251	0.3410
	11				1.0000	0.9998	0.9978	0.9862	0.9313	0.7176
	12					1.0000	1.0000	1.0000	1.0000	1.0000

Tabulated c.d.f. values $F(x) = P(X \leq x)$ for $X \sim \mathscr{B}(12, p)$, $p = 0.1, \ldots, 0.9$.

The appropriate value ≈ 0.9744 can be found in the group corresponding to $n = 12$, in the row corresponding to $x = 3$, and in the column corresponding to $p = 0.1$.

The table can also be used to compute

$$P(X = 3) = P(X \leq 3) - P(X \leq 2) = 0.9744 - 0.8891 \approx 0.0853.$$

- An airline sells 101 tickets for a flight with 100 seats. Each passenger with a ticket is known to have a probability $p = 0.97$ of showing up for their flight. What is the probability of 101 passengers showing up (and the airline being caught overbooking)? Make appropriate assumptions. What if the airline sells 125 tickets?

  Answer: let X be the number of passengers that show up. We want to compute $P(X > 100)$.

  If all passengers show up independently of one another (no families or late bus?), we can model $X \sim \mathscr{B}(101, 0.97)$ and

  $$P(X > 100) = P(X = 101) = \binom{101}{101}(0.97)^{101}(0.03)^0 \approx 0.046.$$

  If the airline sells $n = 125$ tickets, we can model the situation with the binomial distribution $\mathscr{B}(125, 0.97)$, so that

  $$P(X > 100) = 1 - P(X \leq 100) = 1 - \sum_{x=0}^{100} \binom{125}{x}(0.97)^x(0.03)^{125-x}.$$

  This sum is harder to compute directly, but is very nearly 1 (try it in `R`, say).

  Do these results match your intuition?

2.4 Geometric Distributions

Now consider a sequence of Bernoulli trials, with probability p of success at each step. Let the **geometric** random variable X denote the number of steps before the first success occurs.

The probability mass function is given by

$$f(x) = P(X = x) = (1-p)^{x-1}p, \quad x = 1, \ldots,$$

denoted $X \sim \mathrm{Geo}(p)$.

For this random variable, we have

$$\mathrm{E}[X] = \frac{1}{p} \quad \text{and} \quad \mathrm{Var}[X] = \frac{1-p}{p^2}.$$

Examples

- A fair 6–sided die is thrown until it shows a 6. What is the probability that 5 throws are required?

  Answer: If 5 throws are required, we have to compute $P(X = 5)$, where X is geometric Geo(1/6):

  $$P(X = 5) = (1-p)^{5-1}p = (5/6)^4(1/6) \approx 0.0804.$$

- In the example above, how many throws would you expect to need?

  Answer: $\mathrm{E}[X] = \frac{1}{1/6} = 6$.

2.5 Negative Binomial Distribution

Consider now a sequence of Bernoulli trials, with probability p of success at each step. Let the **negative binomial** random variable X denote the number of steps before the rth success occurs.

The probability mass function is given by

$$f(x) = P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r, \quad x = r, \ldots,$$

which we denote by $X \sim \text{NegBin}(p, r)$.

For this random variable, we have

$$\text{E}[X] = \frac{r}{p} \quad \text{and} \quad \text{Var}[X] = \frac{r(1-p)}{p^2}.$$

Example:

- A fair 6—sided die is thrown until it three 6's are rolled. What is the probability that 5 throws are required?

  Answer: If 5 throws are required, we have to compute $P(X = 5)$, where X is geometric NegBin$(1/6, 3)$:

$$\begin{aligned} P(X = 5) &= \binom{5-1}{3-1}(1-p)^{5-3}p^3 \\ &= \binom{4}{2}(5/6)^2(1/6)^3 \approx 0.0193. \end{aligned}$$

- In the example above, how many throws would you expect to need?

  Answer: $\text{E}[X] = \frac{3}{1/6} = 18$.

2.6 Poisson Distributions

Let's say we are counting the number of "changes" that occur in a continuous interval of time or space.[10]

We have a **Poisson process** with rate λ, denoted by $\mathscr{P}(\lambda)$, if:

a) the number of changes occurring in non-overlapping intervals are **independent**;
b) the probability of exactly one change in a short interval of length h is approximately λh, and
c) The probability of 2+ changes in a sufficiently short interval is essentially 0.

Assume that an experiment satisfies the above properties. Let X be the number of changes in a **unit interval** (this could be 1 day, or 15 minutes, or 10 years, etc.).

What is $P(X = x)$, for $x = 0, 1, \ldots$? We can get to the answer by first partition the unit interval into n disjoint sub-intervals of length $1/n$. Then,

1. by condition b), the probability of one change occurring in one of the sub-intervals is approximately λ/n;
2. by condition c), the probability of 2+ changes is ≈ 0, and
3. by condition a), we have a sequence of n Bernoulli trials with probability $p = \lambda/n$.

Therefore,

$$\begin{aligned} f(x) = P(X = x) &\approx \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!} \cdot \underbrace{\frac{n!}{(n-x)!} \cdot \frac{1}{n^x}}_{\text{term 1}} \cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\text{term 2}} \cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^{-x}}_{\text{term 3}}. \end{aligned}$$

Letting $n \to \infty$, we get

$$\begin{aligned} P(X = x) &= \lim_{n\to\infty} \frac{\lambda^x}{x!} \cdot \underbrace{\frac{n!}{(n-x)!} \cdot \frac{1}{n^x}}_{\text{term 1}} \cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\text{term 2}} \cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^{-x}}_{\text{term 3}} \\ &= \frac{\lambda^x}{x!} \cdot 1 \cdot \exp(-\lambda) \cdot 1 = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, \ldots \end{aligned}$$

Let $X \sim \mathscr{P}(\lambda)$. Then it can be shown that

$$\text{E}[X] = \lambda \quad \text{and} \quad \text{Var}[X] = \lambda,$$

that is, the mean and the variance of a Poisson random variable are identical.

Examples:

- A traffic flow is typically modeled by a Poisson distribution. It is known that the traffic flowing through an intersection is 6 cars/minute, on average. What is the probability of no cars entering the intersection in a 30 second period?

  Answer: 6 cars/min = 3 cars/30 sec. Thus $\lambda = 3$, and we need to compute

$$P(X = 0) = \frac{3^0 e^{-3}}{0!} = \frac{e^{-3}}{1} \approx 0.0498.$$

- A hospital needs to schedule night shifts in the maternity ward. It is known that there are 3000 deliveries per year; if these happened randomly round the clock,[11], we would expect 1000 deliveries between the hours of midnight and 8.00 a.m., a time when much of the staff is off-duty.

  It is thus important to ensure that the night shift is sufficiently staffed to allow the maternity ward to cope with the workload on any particular night, or at least, on a high proportion of nights.

[10] Such as # of defects on a production line over a 1 hr period, # of customers that arrive at a teller over a 15 min interval, etc.

[11] Is this a reasonable assumption?

The average number of deliveries per night

$$\lambda = 1000/365.25 \approx 2.74.$$

If the daily number X of night deliveries follows a Poisson process $\mathscr{P}(\lambda)$, we can compute the probability of delivering $x = 0, 1, 2, \ldots$ babies on each night.

Some of the probabilities are:

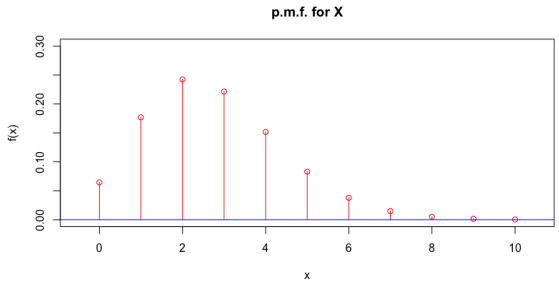
$P(X = x)$	$\lambda^x \cdot \exp(-\lambda)/x!$
$P(X = 0)$	$2.74^0 \cdot \exp(-2.74)/0! = 0.065$
$P(X = 1)$	$2.74^1 \cdot \exp(-2.74)/1! = 0.177$
$P(X = 2)$	$2.74^2 \cdot \exp(-2.74)/2! = 0.242$
$P(X = 3)$	$2.74^3 \cdot \exp(-2.74)/3! = 0.221$
$P(X = 4)$	$2.74^4 \cdot \exp(-2.74)/4! = 0.152$
$P(X = 5)$	$2.74^5 \cdot \exp(-2.74)/5! = 0.083$
$P(X = 6)$	$2.74^6 \cdot \exp(-2.74)/6! = 0.038$
$P(X = 7)$	$2.74^7 \cdot \exp(-2.74)/7! = 0.015$
$P(X = 8)$	$2.74^8 \cdot \exp(-2.74)/8! = 0.005$
$P(X = 9)$	$2.74^9 \cdot \exp(-2.74)/9! = 0.002$
$\vdots$	$\vdots$

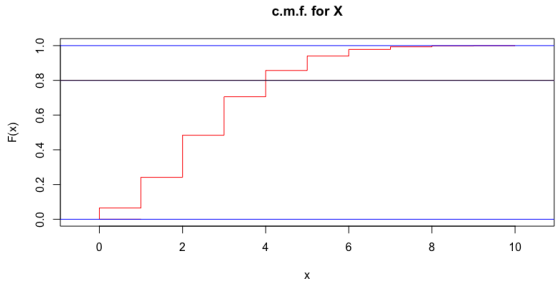
- If the maternity ward wants to prepare for the greatest possible traffic on 80% of the nights, how many deliveries should be expected?

  Answer: we seek an x for which

  $$P(X \leq x-1) \leq 0.80 \leq P(X \leq x):$$

  since $P(X \leq 3) = 0.705$ and $P(X \leq 4) = 0.857$, if they prepare for 4 deliveries a night, they will be ready for the worst on at least 80% of the nights (closer to 85.7%, actually). Note that this is different than asking how many deliveries are expected nightly (namely, $\mathrm{E}[X] = 2.74$).





- On how many nights in the year would 5 or more deliveries be expected?

  Answer: we need to evaluate

  $$365.25 \cdot P(X \geq 5) = 365.25(1 - P(X \leq 4)) = 365.25(1 - 0.857) \approx 52.27.$$

- Over the course of one year, what is the greatest number of deliveries expected on any night?

  Answer: we need to look for largest value of x for which

  $$365.25 \cdot P(X = x) \geq 1.$$

  A few quick computations show that $x = 8$.

2.7 Other Discrete Distributions

Wikipedia [22] lists other common discrete distributions:

- the **Rademacher** distribution, which takes values 1 and -1, each with probability $1/2$;
- the **beta binomial** distribution, which describes the number of successes in a series of independent Bernoulli experiments with heterogeneity in the success probability;
- the **discrete uniform** distribution, where all elements of a finite set are equally likely (balanced coin, unbiased die, first card of a well-shuffled deck, etc.);
- the **hypergeometric** distribution, which describes the number of successes in the first m of a series of n consecutive Bernoulli experiments, if the total number of successes is known;
- the **negative hypergeometric** distribution, which describes the number of attempts needed to get the nth success in a series of Bernoulli experiments;
- the **Poisson binomial** distribution, which describes the number of successes in a series of independent Bernoulli experiments with different success probabilities;
- **Benford's Law**, which describes the frequency of the first digit of many naturally occurring data.
- **Zipf's Law**, which describes the frequency of words in the English language;
- the **beta negative binomial** distribution, which describes the number of failures needed to obtain r successes in a sequence of independent Bernoulli experiments;
- etc.

3. Continuous Distributions

How do we approach probabilities where there there are **uncountably infinitely many possible outcomes**, such as one might encounter if X represents the height of an individual in the population, for instance (e.g., the outcomes reside in a continuous interval)? What is the probability that a randomly selected person is 6 feet tall, say?

3.1 Continuous Random Variables

In the discrete case, the probability mass function

$$f_X(x) = P(X = x)$$

was the main object of interest. In the continuous case, the analogous role is played by the **probability density function** (p.d.f.), still denoted by $f_X(x)$, but now,

$$f_X(x) \neq P(X = x).$$

The **(cumulative) distribution function** (c.d.f.) of any such random variable X is also still defined by

$$F_X(x) = P(X \leq x),$$

viewed as a function of a real variable x; however $P(X \leq x)$ is not simply computed by adding a few terms of the form $P(X = x_i)$ anymore.

Note as well that

$$\lim_{x\to-\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x\to+\infty} F_X(x) = 1.$$

We can describe the **distribution** of the random variable X *via* the following relationship between $f_X(x)$ and $F_X(x)$:

$$f_X(x) = \frac{d}{dx} F_X(x);$$

in the continuous case, probability theory is simply an application of calculus!

Area Under the Curve For any $a < b$, we have

$$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\},$$

so that

$$P(X \leq a) + P(a < X \leq b) = P(X \leq b)$$

and thus

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx \end{aligned}$$

Probability Density Function The **probability density function** (p.d.f.) of a continuous random variable X is an **integrable** function $f_X : X(\mathscr{S}) \to \mathbb{R}$ such that:

- $f_X(x) > 0$ for all $x \in X(\mathscr{S})$ and $\lim_{x\to\pm\infty} f_X(x) = 0$;
- $\int_{\mathscr{S}} f_X(x)\,dx = 1$;
- for any event $A = (a, b) = \{X | a < X < b\}$,

$$P(A) = P((a, b)) = \int_a^b f_X(x)\,dx,$$

and the **cumulative distribution function** (c.d.f.) F_X is given by

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

Unlike discrete distributions, the absence or presence of endpoints does not affect the probability computations for continuous distributions: for any a, b,

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b),$$

all taking the value

$$F_X(b) - F_X(a) = \int_a^b f(x)\,dx.$$

Furthermore, for any x,

$$P(x > X) = 1 - P(X \leq x) = 1 - F_X(x) = 1 - \int_{-\infty}^{x} f_X(t)\,dt;$$

and for any a,

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f_X(x)\,dx = 0.$$

That last result explains why it is pointless to speak of the probability of a random variable taking on a specific value in the continuous case; rather, we are interested in **ranges** of values.

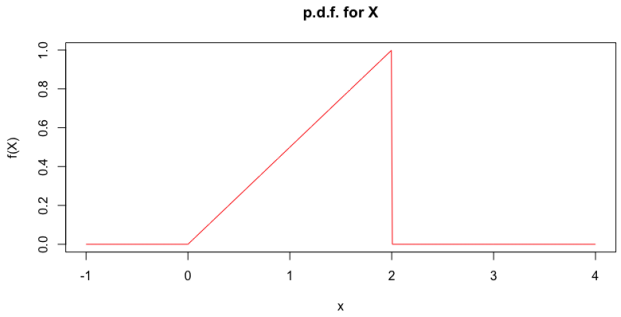
Examples

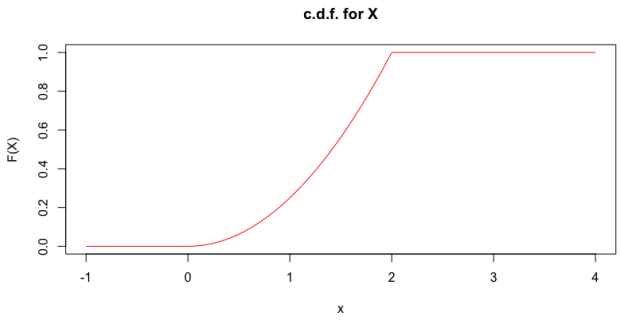
- Assume that X has the following p.d.f.:

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x/2 & \text{if } 0 \leq x \leq 2 \\ 0 & \text{if } x > 2 \end{cases}$$

Note that $\int_0^2 f(x)\,dx = 1$. The corresponding c.d.f. is given by:

$$\begin{aligned} F_X(x) = P(X \leq x) &= \int_{-\infty}^{x} f_X(t)\,dt \\ &= \begin{cases} 0 & \text{if } x < 0 \\ 1/2 \cdot \int_0^x t\,dt = x^2/4 & \text{if } 0 < x < 2 \\ 1 & \text{if } x \geq 2 \end{cases} \end{aligned}$$



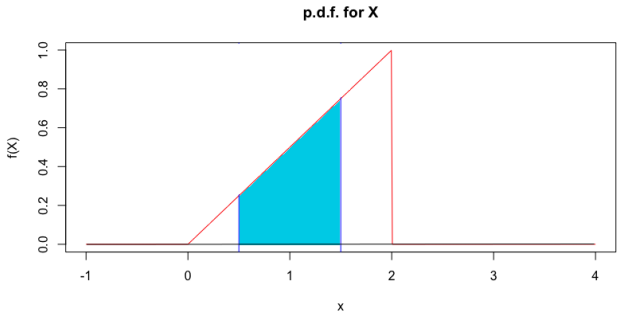


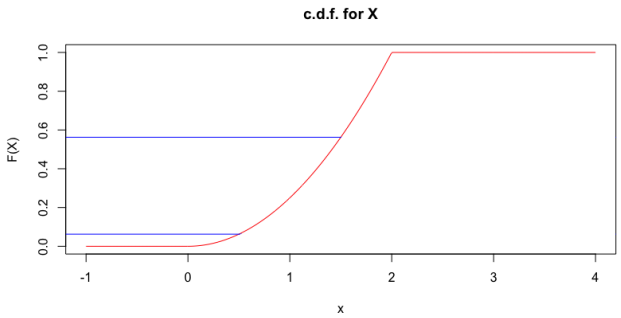
- What is the probability of the event

$$A = \{X | 0.5 < X < 1.5\}?$$

Answer: we need to evaluate

$$P(A) = P(0.5 < X < 1.5) = F_X(1.5) - F_X(0.5)$$
$$= \frac{(1.5)^2}{4} - \frac{(0.5)^2}{4} = \frac{1}{2}.$$





- What is the probability of the event $B = \{X | X = 1\}$?

Answer: we need to evaluate

$$P(B) = P(X = 1) = P(1 \le X \le 1) = F_X(1) - F_X(1) = 0.$$

This is not unexpected: even though $f_X(1) = 0.5 \neq 0$, $P(X = 1) = 0$, as we say earlier.

- Assume that, for $\lambda > 0$, X has the following p.d.f.:

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \ge 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Verify that f_X is a p.d.f. for all $\lambda > 0$, and compute the probability that $X > 10.2$.

Answer: that f_X is a p.d.f. is obvious; the only work goes into showing that

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^{\infty} \lambda \exp(-\lambda x)\,dx$$
$$= \lim_{b\to\infty} \int_0^b \lambda \exp(-\lambda x)\,dx$$
$$= \lim_{b\to\infty} \lambda \left[\frac{\exp(-\lambda x)}{-\lambda}\right]_0^b = \lim_{b\to\infty} [-\exp(-\lambda x)]_0^b$$
$$= \lim_{b\to\infty} [-\exp(-\lambda b) + \exp(0)] = 1.$$

The corresponding c.d.f. is given by:

$$F_X(x;\lambda) = P_\lambda(X \le x) = \int_{-\infty}^{x} f_X(t)\,dt$$
$$= \begin{cases} 0 & \text{if } x < 0 \\ \lambda \int_0^x \exp(-\lambda t)\,dt & \text{if } x \ge 0 \end{cases}$$
$$= \begin{cases} 0 & \text{if } x < 0 \\ [-\exp(-\lambda t)]_0^x & \text{if } x \ge 0 \end{cases}$$
$$= \begin{cases} 0 & \text{if } x < 0 \\ 1 - \exp(-\lambda x) & \text{if } x \ge 0 \end{cases}$$

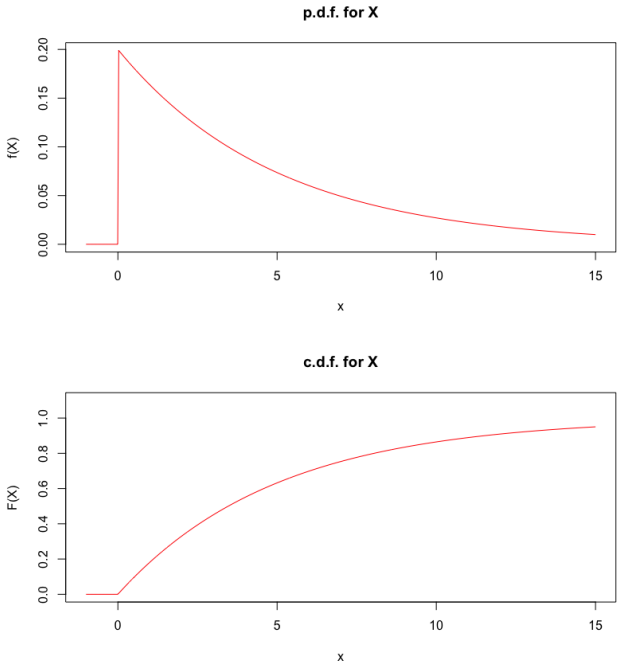
Then

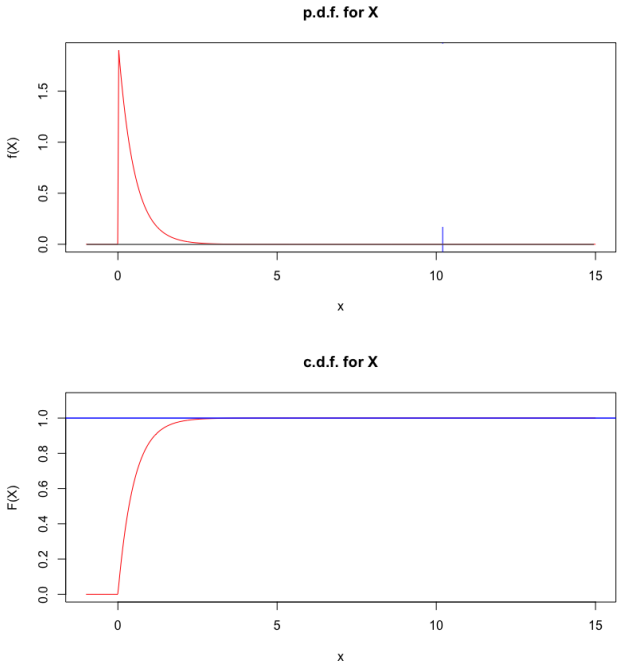
$$P_\lambda(X > 10.2) = 1 - F_X(10.2;\lambda) = 1 - [1 - \exp(-10.2\lambda)]$$
$$= \exp(-10.2\lambda)$$

is a function of the **distribution parameter** λ itself:

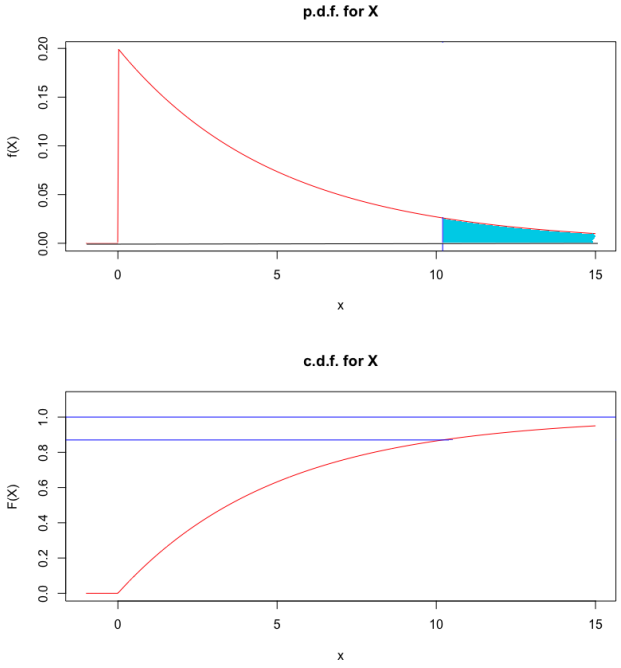
λ	$P_\lambda(X > 10.2)$
0.002	0.9798
0.02	0.8155
0.2	0.1300
2	1.38×10^{-9}
20	2.54×10^{-89}
200	0 (for all intents and purposes)

For $\lambda = 0.2$, for instance, the p.d.f. and c.d.f. are:





and the probability that $X > 10.2$ is the area (to ∞) in blue, below.



For $\lambda = 2$, the probability is so small (1.38×10^{-9}) that it does not appear in the p.d.f. in the next column over.

Note that in all cases, the **shape** of the p.d.f. and the c.d.f are the same (the spike when $\lambda = 2$ is much higher than that when $\lambda = 0.2$ – why must that be the case?). This is not a general property of distributions, however, but a property of this specific family of distributions.

3.2 Expectation of a Continuous Random Variable

For a continuous random variable X with p.d.f. $f_X(x)$, the **expectation** of X is defined as

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx\,.$$

For any function $h(X)$, we can also define

$$\mathrm{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)\, dx\,.$$

Examples:

- Find $\mathrm{E}[X]$ and $\mathrm{E}[X^2]$ in the example on p. 17.

  Answer: we need to evaluate

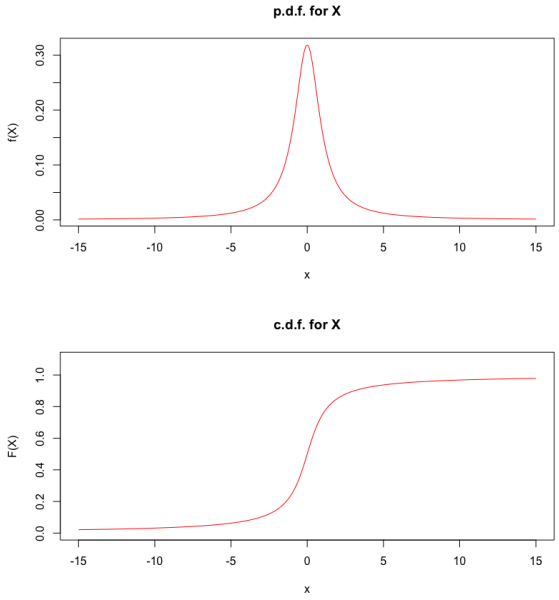
  $$\begin{aligned}\mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(X)\, dx = \int_0^2 x f_X(x)\, dx \\ &= \int_0^2 \frac{x^2}{2}\, dx = \left[\frac{x^3}{6}\right]_{x=0}^{x=2} = \frac{4}{3}; \\ \mathrm{E}[X^2] &= \int_0^2 \frac{x^3}{2}\, dx = 2.\end{aligned}$$

- Note that the **expectation need not exist**! Compute the expectation of the random variable X with p.d.f.

  $$f_X(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

  Answer: let's verify that $f_X(x)$ is indeed a p.d.f.:

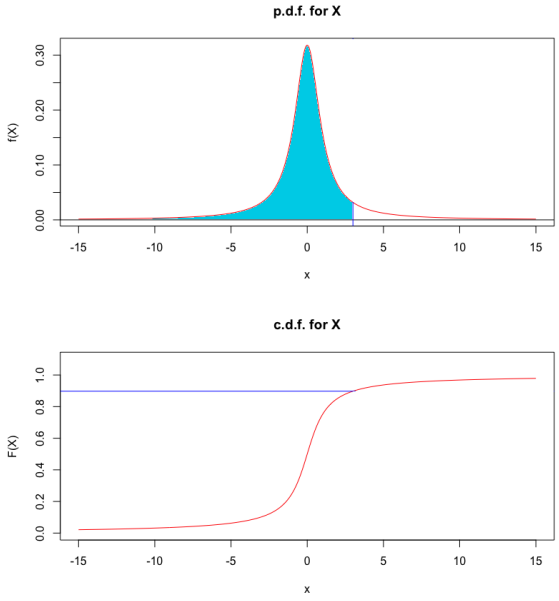
  $$\begin{aligned}\int_{-\infty}^{\infty} f_X(x)\, dx &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{1+x^2}\, dx \\ &= \frac{1}{\pi}[\arctan(x)]_{-\infty}^{\infty} = \frac{1}{\pi}\left[\frac{\pi}{2} + \frac{\pi}{2}\right] = 1.\end{aligned}$$



We can also easily see that

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt$$
$$= \frac{1}{\pi}\int_{-\infty}^{x} \frac{1}{1+t^2}\,dt = \frac{1}{\pi}\arctan(x) + \frac{1}{2}.$$

For instance, $P(X \leq 3) = \frac{1}{\pi}\arctan(3) + \frac{1}{2}$, say.



The expectation of X is

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\,dx.$$

If this improper integral exists, then it needs to be equal **both** to

$$\underbrace{\int_{-\infty}^{0} \frac{x}{\pi(1+x^2)}\,dx + \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,dx}_{\text{candidate 1}}$$

and to the Cauchy principal value

$$\underbrace{\lim_{a\to\infty} \int_{-a}^{a} \frac{x}{\pi(1+x^2)}\,dx}_{\text{candidate 2}}.$$

But it is straightforward to find an antiderivative of $\frac{x}{\pi(1+x^2)}$. Set $u = 1+x^2$. Then $du = 2x\,dx$ and $x\,dx = \frac{du}{2}$, and we obtain

$$\int \frac{x}{\pi(1+x^2)}\,dx = \frac{1}{2\pi}\int u\,du = \frac{1}{2\pi}\ln|u| = \frac{1}{2\pi}\ln(1+x^2).$$

Then the candidate 2 integral reduces to

$$\lim_{a\to\infty}\left[\frac{\ln(1+x^2)}{2\pi}\right]_{-a}^{a} = \lim_{a\to\infty}\left[\frac{\ln(1+a^2)}{2\pi} - \frac{\ln(1+(-a)^2)}{2\pi}\right]$$
$$= \lim_{a\to\infty} 0 = 0;$$

while the candidate 1 integral reduces to

$$\left[\frac{\ln(1+x^2)}{2\pi}\right]_{-\infty}^{0} + \left[\frac{\ln(1+x^2)}{2\pi}\right]_{0}^{\infty} = 0 - (\infty) + \infty - 0 = \infty - \infty$$

which is **undefined**. Thus $\mathrm{E}[X]$ cannot not exist, as it would have to be both equal to 0 and be undefined simultaneously.

Mean and Variance In a similar way to the discrete case, the **mean** of X is defined to be $\mathrm{E}[X]$, and the **variance** and **standard deviation** of X are, as before,

$$\mathrm{Var}[X] \overset{\text{def}}{=} \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] = \mathrm{E}[X^2] - \mathrm{E}^2[X],$$
$$\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}.$$

As in the discrete case, if X, Y are continuous random variables, and $a, b \in \mathbb{R}$, then

$$\mathrm{E}[aY + bX] = a\mathrm{E}[Y] + b\mathrm{E}[X]$$
$$\mathrm{Var}[a + bX] = b^2\mathrm{Var}[X]$$
$$\mathrm{SD}[a + bX] = |b|\,\mathrm{SD}[X]$$

The interpretations of the mean as a measure of **centrality** and of the variance as a measure of **dispersion** are unchanged in the continuous case.

For the time being, however, we cannot easily compute the variance of a sum $X + Y$, unless X and Y are **independent** random variables, in which case

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

3.3 Normal Distributions

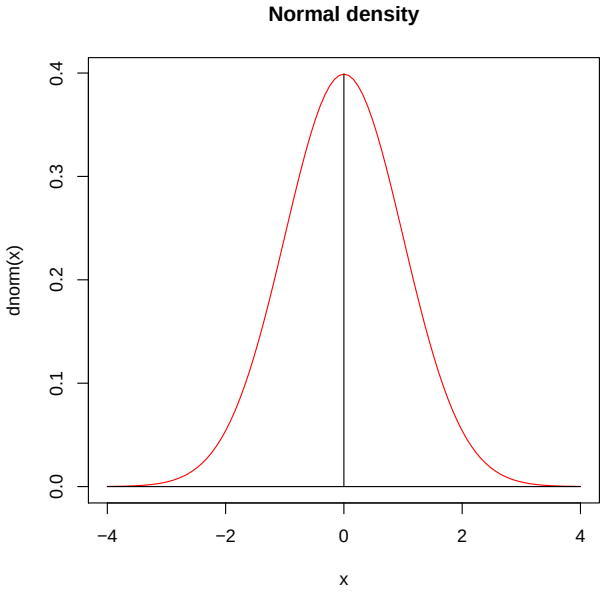
A **very** important example of a continuous distribution is that provided by the special probability distribution function

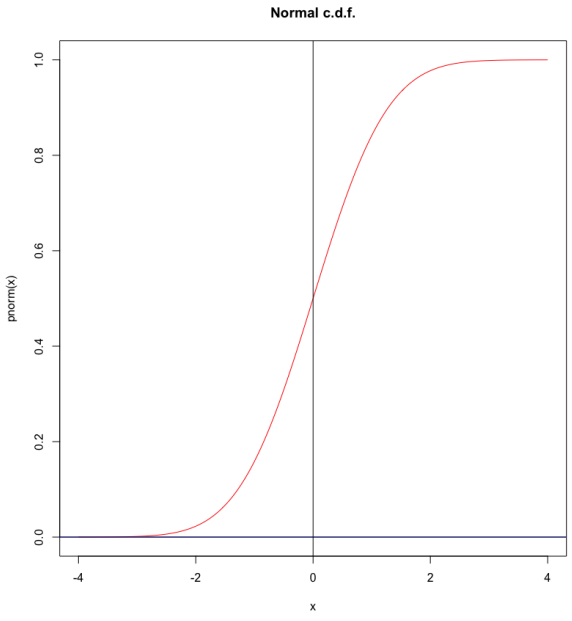
$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

The corresponding cumulative distribution function is denoted by

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \phi(t)\, dt\,.$$

A random variable Z with this c.d.f. is said to have a **standard normal distribution**, denoted by $Z \sim \mathcal{N}(0,1)$.





Standard Normal Random Variable The expectation and variance of $Z \sim \mathcal{N}(0,1)$ are

$$\begin{aligned}
\mathrm{E}[Z] &= \int_{-\infty}^{\infty} z\,\phi(z)\,dz = \int_{-\infty}^{\infty} z\,\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,dz = 0,\\
\mathrm{Var}[Z] &= \int_{-\infty}^{\infty} z^2\,\phi(z)\,dz = 1,\\
\mathrm{SD}[Z] &= \sqrt{\mathrm{Var}[Z]} = \sqrt{1} = 1.
\end{aligned}$$

Other quantities of interest include:

$$\begin{aligned}
&\Phi(0) = P(Z \le 0) = \frac{1}{2}, \quad \Phi(-\infty) = 0, \quad \Phi(\infty) = 1,\\
&\Phi(1) = P(Z \le 1) \approx 0.8413, \quad \text{etc.}
\end{aligned}$$

Normal Random Variables Let $\sigma > 0$ and $\mu \in \mathbb{R}$.

If $Z \sim \mathcal{N}(0,1)$ and $X = \mu + \sigma Z$, then

$$\frac{X-\mu}{\sigma} = Z \sim \mathcal{N}(0,1).$$

Thus, the c.d.f. of X is given by

$$\begin{aligned}
F_X(x) &= P(X \le x)\\
&= P(\mu + \sigma Z \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right)\\
&= \Phi\left(\frac{x-\mu}{\sigma}\right);
\end{aligned}$$

its p.d.f. must then be

$$\begin{aligned}
f_X(x) &= \frac{d}{dx} F_X(x)\\
&= \frac{d}{dx}\Phi\left(\frac{x-\mu}{\sigma}\right)\\
&= \frac{1}{\sigma}\,\phi\left(\frac{x-\mu}{\sigma}\right).
\end{aligned}$$

Any random variable X with this c.d.f./p.d.f. satisfies

$$\begin{aligned}
\mathrm{E}[X] &= \mu + \sigma \mathrm{E}[Z] = \mu,\\
\mathrm{Var}[X] &= \sigma^2 \mathrm{Var}[Z] = \sigma^2,\\
\mathrm{SD}[X] &= \sigma
\end{aligned}$$

and is said to be **normal with mean μ and variance σ^2**, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$.

As it happens, every general normal X can be obtained by a linear transformation of the standard normal Z.

Traditionally, probability computations for normal distributions are done with tables which compile values of the standard normal distribution c.d.f., such as the one found in [21] (see here for a preview).

With the advent of freely-available **statistical software**, the need for tabulated values had decreased.[12]

In R, the standard normal c.d.f. $F_Z(z) = P(Z \le z)$ can be computed with the function `pnorm(z)` – for instance, `pnorm(0)=0.5`. (In the example below, whenever $P(Z \le a)$ is evaluated for some a, the value is found either by consulting a table or using `pnorm`.)

Examples

- Let Z represent the standard normal random variable. Then:
  a) $P(Z \le 0.5) = 0.6915$
  b) $P(Z < -0.3) = 0.3821$
  c) $P(Z > 0.5) = 1 - P(Z \le 0.5) = 1 - 0.6915 = 0.3085$
  d) $P(0.1 < Z < 0.3) = P(Z < 0.3) - P(Z < 0.1) = 0.6179 - 0.5398 = 0.0781$
  e) $P(-1.2 < Z < 0.3) = P(Z < 0.3) - P(Z < -1.2) = 0.5028$

- Suppose that the waiting time (in minutes) in a coffee shop at 9am is normally distributed with mean 5 and standard deviation 0.5.[13] What is the probability that the waiting time for a customer is at most 6 minutes?

  Answer: let X denote the waiting time.

  Then $X \sim \mathcal{N}(5, 0.5^2)$ and the **standardised random variable** is a standard normal:

$$Z = \frac{X-5}{0.5} \sim \mathcal{N}(0,1).$$

  The desired probability is

$$\begin{aligned} P(X \le 6) &= P\left(\frac{X-5}{0.5} \le \frac{6-5}{0.5}\right) \\ &= P\left(Z \le \frac{6-5}{0.5}\right) = \Phi\left(\frac{6-5}{0.5}\right) \\ &= \Phi(2) = P(Z \le 2) \approx 0.9772. \end{aligned}$$

- Suppose that bottles of beer are filled in such a way that the actual volume of the liquid content (in mL) varies randomly according to a normal distribution with $\mu = 376.1$ and $\sigma = 0.4$.[14] What is the probability that the volume in any randomly selected bottle is less than 375mL?

  Answer: let X denote the volume of the liquid in the bottle. Then

$$X \sim \mathcal{N}(376.1, 0.4^2) \implies Z = \frac{X-376.1}{0.4} \sim \mathcal{N}(0,1).$$

  The desired probability is thus

$$\begin{aligned} P(X < 375) &= P\left(\frac{X-376.1}{0.4} < \frac{375-376.1}{0.4}\right) \\ &= P\left(Z < \frac{-1.1}{0.4}\right) \\ &= P(Z \le -2.75) = \Phi(-2.75) \approx 0.003. \end{aligned}$$

- If $Z \sim \mathcal{N}(0,1)$, for which values a, b and c do:
  a) $P(Z \le a) = 0.95$;
  b) $P(|Z| \le b) = P(-b \le Z \le b) = 0.99$;
  c) $P(|Z| \ge c) = 0.01$.

  Answer:

  a) From the table (or R) we see that

$$P(Z \le 1.64) \approx 0.9495,\ P(Z \le 1.65) \approx 0.9505.$$

  Clearly we must have $1.64 < a < 1.65$; a linear interpolation provides a decent guess at $a \approx 1.645$.[15].

  b) Note that

$$P(-b \le Z \le b) = P(Z \le b) - P(Z < -b)$$

  However the p.d.f. $\phi(z)$ is symmetric about $z = 0$, which means that

$$P(Z < -b) = P(Z > b) = 1 - P(Z \le b),$$

  and so that

$$\begin{aligned} P(-b \le Z \le b) &= P(Z \le b) - [1 - P(Z \le b)] \\ &= 2P(Z \le b) - 1 \end{aligned}$$

  In the question, $P(-b \le Z \le b) = 0.99$, so that

$$2P(Z \le b) - 1 = 0.99 \implies P(Z \le b) = \frac{1+0.99}{2} = 0.995.$$

  Consulting the table we see that

$$P(Z \le 2.57) \approx 0.9949,\ P(Z \le 2.58) \approx 0.9951;$$

  a linear interpolation suggests that $b \approx 2.575$.

  c) Note that $\{|Z| \ge c\} = \{|Z| < c\}^c$, so we need to find c such that

$$P(|Z| < c) = 1 - P(|Z| \ge c) = 0.99.$$

  But this is equivalent to

$$P(-c < Z < c) = P(-c \le Z \le c) = 0.99$$

  as $|x| < y \Leftrightarrow -y < x < y$, and $P(Z = c) = 0$ for all c. This problem was solved in part b); set $c \approx 2.575$.

[12] Although it would still be a good idea to learn how to read/use them.

[13] In theory, this cannot be the true model as this would imply that some of the wait times could be negative, but it may nevertheless be an acceptable assumption in practice.

[14] The statement from the previous footnote applies here as well – we will assume that this is understood from this point onward.

[15] This level of precision is usually not necessary – it is often sufficient to simply present the interval estimate: $a \in (1.64, 1.65)$

3.4 Exponential Distributions

Assume that cars arrive according to a **Poisson process with rate** λ, that is, the number of cars arriving within a fixed unit time period is a Poisson random variable with parameter λ.

Over a period of time x, we would then expect the number of arrivals N to follow a Poisson process with parameter λx. Let X be the wait time to the first car arrival. Then

$$P(X > x) = 1 - P(X \le x) = P(N = 0) = \exp(-\lambda x).$$

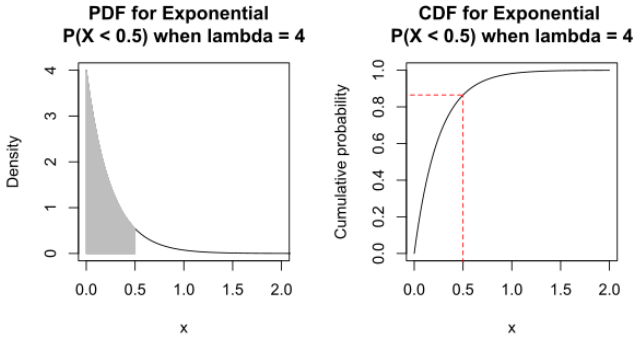
We say that X follows an **exponential distribution** $\text{Exp}(\lambda)$:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-\lambda x} & \text{for } 0 \le x \end{cases}$$

$$f_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \lambda e^{-\lambda x} & \text{for } 0 \le x \end{cases}$$

Note that $f_X(x) = F_X'(x)$ for all x.

If $X \sim \text{Exp}(4)$, then $P(X < 0.5) = F_X(0.5) = 1 - e^{-4(0.5)} \approx 0.865$ is the area of the shaded region below:



Properties If $X \sim \text{Exp}(\lambda)$, then

- $\mu = \text{E}[X] = 1/\lambda$, since

$$\begin{aligned} \mu &= \int_0^\infty x\lambda e^{-\lambda x}\,dx = \left[-\frac{\lambda x + 1}{\lambda} e^{-\lambda x}\right]_0^\infty \\ &= \left[0 + \frac{\lambda(0)+1}{\lambda} e^{-0}\right] \\ &= \frac{1}{\lambda}; \end{aligned}$$

- $\sigma^2 = \text{Var}[X] = 1/\lambda^2$, since

$$\begin{aligned} \sigma^2 &= \int_0^\infty (x - \text{E}[X])^2\,\lambda e^{-\lambda x}\,dx \\ &= \int_0^\infty \left(x - \frac{1}{\lambda}\right)^2 \lambda e^{-\lambda x}\,dx \\ &= \left[-\frac{\lambda^2 x^2 + 1}{\lambda^2} e^{-\lambda x}\right]_0^\infty \\ &= \left[0 + \frac{\lambda^2(0)^2 + 1}{\lambda^2} e^{-0}\right] \\ &= \frac{1}{\lambda^2}; \end{aligned}$$

- and $P(X > s + t \mid X > t) = P(X > s)$, for all $s, t > 0$, since

$$\begin{aligned} P(X > s+t \mid X > t) &= \frac{P(X > s+t \text{ and } X > t)}{P(X > t)} \\ &= \frac{P(X > s+t)}{P(X > t)} = \frac{1 - F_X(s+t)}{1 - F_X(t)} \\ &= \frac{\exp(-\lambda(s+t))}{\exp(-\lambda t)} \\ &= \exp(-\lambda s) = P(X > s) \end{aligned}$$

(we say that exponential distributions are **memory-less**).

In a sense, $\text{Exp}(\lambda)$ is the continuous analogue to the **geometric** distribution $\text{Geo}(p)$.

Example: the lifetime of a certain type of light bulb follows an exponential distribution whose mean is 100 hours (i.e. $\lambda = 1/100$).

- What is the probability that a light bulb will last at least 100 hours?

  Answer: Since $X \sim \text{Exp}(1/100)$, we have

  $$P(X > 100) = 1 - P(X \le 100) = \exp(-100/100) \approx 0.37.$$

- Given that a light bulb has already been burning for 100 hours, what is the probability that it will last at least 100 hours more?

  Answer: we seek $P(X > 200 \mid X > 100)$. By the memory-less property,

  $$P(X > 200 \mid X > 100) = P(X > 200 - 100) = P(X > 100) \approx 0.37.$$

- The manufacturer wants to guarantee that their light bulbs will last at least t hours. What should t be in order to ensure that 90% of the light bulbs will last longer than t hours?

  Answer: we need to find t such that $P(X > t) = 0.9$. In other words, we are looking for t such that

  $$0.9 = P(X > t) = 1 - P(X \le t) = 1 - F_X(t) = e^{-0.01t},$$

  that is,

  $$\ln 0.9 = -0.01t \Longrightarrow t = -100 \ln 0.9 \approx 10.5 \text{ hours}.$$

3.5 Gamma Distributions

Assume that cars arrive according to a Poisson process with rate λ. Recall that if X is the time to the first car arrival, then $X \sim \text{Exp}(\lambda)$.

If Y is the wait time to the rth arrival, then Y follows a **Gamma distribution** with parameters λ and r, denoted $Y \sim \Gamma(\lambda, r)$, for which the p.d.f. is

$$f_Y(y) = \begin{cases} 0 & \text{for } y < 0 \\ \frac{y^{r-1}}{(r-1)!}\lambda^r e^{-\lambda y} & \text{for } y \ge 0 \end{cases}$$

The c.d.f. $F_Y(y)$ exists (it is the area under f_Y from 0 to y), but it cannot be expressed with elementary functions.

We can show that

$$\mu = \mathrm{E}[Y] = \frac{r}{\lambda} \quad \text{and} \quad \sigma^2 = \mathrm{Var}[Y] = \frac{r}{\lambda^2}.$$

Examples

- Suppose that an average of 30 customers per hour arrive at a shop in accordance with a Poisson process, that is to say, $\lambda = 1/2$ customers arrive on average every minute. What is the probability that the shopkeeper will wait more than 5 minutes before both of the first two customers arrive?

  Answer: let Y denote the wait time in minutes until the second customer arrives. Then $Y \sim \Gamma(1/2, 2)$ and

$$\begin{aligned} P(Y > 5) &= \int_5^\infty \frac{y^{2-1}}{(2-1)!}(1/2)^2 e^{-y/2}\, dy \\ &= \int_5^\infty \frac{y e^{-y/2}}{4}\, dy \\ &= \frac{1}{4}\left[-2y e^{-y/2} - 4e^{-y/2}\right]_5^\infty \\ &= \frac{7}{2}e^{-5/2} \approx 0.287. \end{aligned}$$

- Telephone calls arrive at a switchboard at a mean rate of $\lambda = 2$ per minute, according to a Poisson process. Let Y be the waiting time until the 5th call arrives. What is the p.d.f., the mean, and the variance of Y?

  Answer: we have

$$\begin{aligned} f_Y(y) &= \frac{2^5 y^4}{4!} e^{-2y}, \text{ for } 0 \le y < \infty, \\ \mathrm{E}[Y] &= \frac{5}{2}, \quad \mathrm{Var}[Y] = \frac{5}{4}. \end{aligned}$$

The Gamma distribution can be extended to cases where $r > 0$ is not an integer by replacing $(r-1)!$ by

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t}\, dt.$$

The exponential and the χ^2 distributions (we will discuss the latter later) are special cases of the Gamma distribution: $\mathrm{Exp}(\lambda) = \Gamma(\lambda, 1)$ and $\chi^2(r) = \Gamma(1/2, r)$.

3.6 Normal Approximation of the Binomial Distribution

If $X \sim \mathscr{B}(n,p)$ then we may interpret X as a sum of **independent and identically distributed** random variables

$$X = I_1 + I_2 + \cdots + I_n \text{ where each } I_i \sim \mathscr{B}(1,p).$$

Thus, according to the **Central Limit Theorem** (we'll have more to say on this topic in Section 6.2), for large n we have

$$\frac{X - np}{\sqrt{np(1-p)}} \overset{\text{approx}}{\sim} \mathscr{N}(0,1);$$

for large n if $X \overset{\text{exact}}{\sim} \mathscr{B}(n,p)$ then $X \overset{\text{approx}}{\sim} \mathscr{N}(np, np(1-p))$.

Normal Approximation with Continuity Correction When $X \sim \mathscr{B}(n,p)$, we know that $\mathrm{E}[X] = np$ and $\mathrm{Var}[X] = np(1-p)$. If n is large, we may approximate X by a normal random variable in the following way:

$$P(X \le x) = P(X < x + 0.5) = P\left(Z < \frac{x - np + 0.5}{\sqrt{np(1-p)}}\right)$$

and

$$P(X \ge x) = P(X > x - 0.5) = P\left(Z > \frac{x - np - 0.5}{\sqrt{np(1-p)}}\right).$$

The continuity correction terms are the corresponding ± 0.5 in the expressions (they are required).

Example: suppose $X \sim \mathscr{B}(36, 0.5)$. Provide a normal approximation to the probability $P(X \le 12)$.[16]

Answer: the expectation and the variance of a binomial r.v. are known:

$$\mathrm{E}[X] = 36(0.5) = 18 \quad \text{and} \quad \mathrm{Var}[X] = 36(0.5)(1-0.5) = 9,$$

and so

$$\begin{aligned} P(X \le 12) &= P\left(\frac{X - 18}{3} \le \frac{12 - 18 + 0.5}{3}\right) \\ &\overset{\text{norm.approx'n}}{\approx} \Phi(-1.83) \overset{\text{table}}{\approx} 0.033. \end{aligned}$$

Computing Binomial Probabilities There are thus at least four ways of computing (or approximating) binomial probabilities:

- using the exact formula – if $X \sim \mathscr{B}(n,p)$ then for each $x = 0, 1, \ldots, n$, $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$;
- using tables: if $n \le 15$ and p is one of $0.1, \ldots, 0.9$, then the corresponding c.d.f. can be found in many textbook (we must first express the desired probability in terms of the c.d.f. $P(X \le x)$), such as in

$$\begin{aligned} &P(X < 3) = P(X \le 2); \\ &P(X = 7) = P(X \le 7) - P(X \le 6); \\ &P(X > 7) = 1 - P(X \le 7); \\ &P(X \ge 5) = 1 - P(X \le 4), \text{ etc.} \end{aligned}$$

[16] The binomial probabilities are not typically available in textbooks (or online) for $n = 36$, although they could be computed directly in `R`, such as with `pbinom(12,26,0.5)=0.0326`.

- using statistical software (`pbinom()` in R, say), and
- using the normal approximation when np and $n(1-p)$ are both ≥ 5:

$$P(X \leq x) \approx \Phi\left(\frac{x+0.5-np}{\sqrt{np(1-p)}}\right)$$

$$P(X \geq x) \approx 1-\Phi\left(\frac{x-0.5-np}{\sqrt{np(1-p)}}\right).$$

3.7 Other Continuous Distributions

Other common continuous distributions are listed in [22]:

- the **Beta** distribution, a family of 2-parameter distributions with one mode and which is useful to estimate success probabilities (special cases: uniform, arcsine, PERT distributions);
- the **logit-normal** distribution on $(0,1)$, which is used to model proportions;
- the **Kumaraswamy** distribution, which is used in simulations in lieu of the Beta distribution (as it has a closed form c.d.f.);
- the **triangular** distribution, which is typically used as a subjective description of a population for which there is only limited sample data (it is based on a knowledge of the minimum and maximum and a guess of the mode);
- the **chi-squared** distribution, which is the sum of the squares of n independent normal random variables, is used in goodness-of-fit tests in statistics;
- the $F-$distribution, which is the ratio of two chi-squared random variables, used in the analysis of variance;
- the **Erlang** distribution is the distribution of the sum of k independent and identically distributed exponential random variables, and it is used in queueing models (it is a special case of the Gammma distribution);
- the **Pareto** distribution, which is used to describe financial data and critical behavior;
- **Student's T statistic**, which arise when estimating the mean of a normally-distributed population in situations where the sample size is small and the population's standard deviation is unknown;
- the **logistic** distribution, whose cumulative distribution function is the logistic function;
- the **log-normal** distribution, which describing variables that are the product of many small independent positive variables;
- etc.

4. Joint Distributions

Let X, Y be two continuous random variables. The **joint probability distribution function** (joint p.d.f.) of X, Y is a function $f(x,y)$ satisfying:

1. $f(x,y) \geq 0$, for all x, y;
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$, and
3. $P(A) = \iint_A f(x,y)\,dx\,dy$, where $A \subseteq \mathbb{R}^2$.

For a discrete variable, the properties are the same, except that we replace integrals by sums, and we add a property to the effect that $f(x,y) \leq 1$ for all x, y.

Property 3 implies that $P(A)$ is the **volume** of the solid over the region A in the xy plane bounded by the surface $z = f(x,y)$.

Examples:

- Roll a pair of unbiased dice. For each of the 36 possible outcomes, let X denote the smaller roll, and Y the larger roll (taken from [7]).

  a) How many outcomes correspond to the event $A = \{(X=2, Y=3)\}$?

  Answer: the rolls $(3,2)$ and $(2,3)$ both give rise to event A.

  b) What is $P(A)$?

  Answer: there are 36 possible outcomes, so $P(A) = \frac{2}{36} \approx 0.0556$.

  c) What is the joint p.m.f. of X, Y?

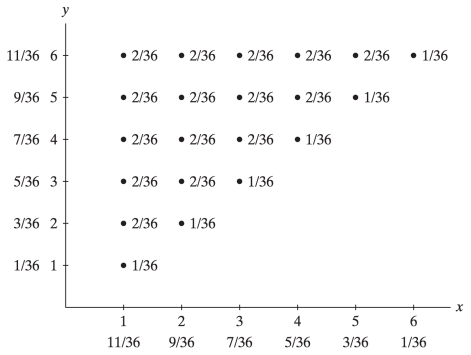
  Answer: only one outcome, $(X=a, Y=a)$, gives rise to the event $\{X=Y=a\}$. For every other event $\{X \neq Y\}$, two outcomes do the trick: (X,Y) and (Y,X). The joint p.m.f. is thus

$$f(x,y) = \begin{cases} 1/36 & 1 \leq x = y \leq 6 \\ 2/36 & 1 \leq x < y \leq 6 \end{cases}$$

  The first property is automatically satisfied, as is the third (by construction). There are only 6 outcomes for which $X = Y$, all the remaining outcomes (of which there are 15) have $X < Y$.

  Thus,

$$\sum_{x=1}^{6}\sum_{y=x}^{6} f(x,y) = 6 \cdot \frac{1}{36} + 15 \cdot \frac{2}{36} = 1.$$



d) Compute $P(X=a)$ and $P(Y=b)$, for $a,b=1,\dots,6$.

Answer: for every $a=1,\dots,6$, $\{X=a\}$ corresponds to the following union of events:

$$\{X=a,Y=a\}\cup\{X=a,Y=a+1\}\cup\cdots \cdots\cup\{X=a,Y=6\}.$$

These events are mutually exclusive, so that

$$\begin{aligned}P(X=a)&=\sum_{y=a}^{6}P(\{X=a,Y=y\})\\&=\frac{1}{36}+\sum_{y=a+1}^{6}\frac{2}{36}\\&=\frac{1}{36}+\frac{2(6-a)}{36},\quad a=1,\dots,6.\end{aligned}$$

Similarly, we get

$$P(Y=b)=\frac{1}{36}+\frac{2(b-6)}{36},\ b=1,\dots,6.$$

These **marginal probabilities** can be found in the margins of the p.m.f.

e) Compute $P(X=3\mid Y>3)$, $P(Y\le 3\mid X\ge 4)$.

Answer: the notation suggests how to compute these **conditional probabilities**:

$$P(X=3\mid Y>3)=\frac{P(X=3\cap Y>3)}{P(Y>3)}$$

The region corresponding to $P(Y>3)=\frac{27}{36}$ is shaded in red (see image at the top of the following column); the region corresponding to $P(X=3)=\frac{7}{36}$ is shaded in blue.

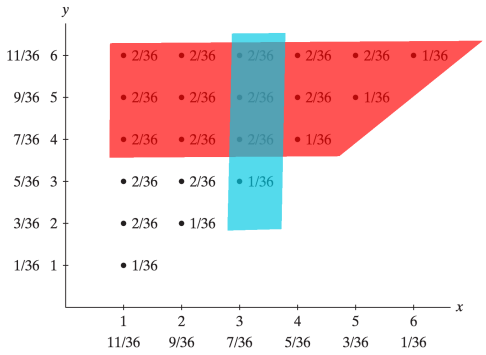
The region corresponding to

$$P(X=3\cap Y>3)=\frac{6}{36}$$

is the intersection of the regions:

$$P(X=3\mid Y>3)=\frac{6/36}{27/36}=\frac{6}{27}\approx 0.2222.$$

As $P(Y\le 3\cap X\ge 4)=0$, $P(Y\le 3|X\ge 4)=0$.



f) Are X and Y independent?

Answer: why didn't we simply use the multiplicative rule to compute

$$P(X=3\cap Y>3)=P(X=3)P(Y>3)?$$

It's because X and Y are **not independent**, that is, it is not always the case that

$$P(X=x,Y=y)=P(X=x)P(Y=y)$$

for all allowable x,y.

As it is, $P(X=1,Y=1)=\frac{1}{36}$, but

$$P(X=1)P(Y=1)=\frac{11}{36}\cdot\frac{1}{36}\neq\frac{1}{36},$$

so X and Y are **dependent** (this is often the case when the domain of the joint p.d.f./p.m.f. is not rectangular).

- There are 8 similar chips in a bowl: three marked $(0,0)$, two marked $(1,0)$, two marked $(0,1)$ and one marked $(1,1)$. A player selects a chip at random and is given the sum of the two coordinates, in dollars (taken from [7]).

a) What is the joint probability mass function of X_1, and X_2?

Answer: let X_1 and X_2 represent the coordinates; we have

$$f(x_1,x_2)=\frac{3-x_1-x_2}{8},\quad x_1,x_2=0,1.$$

a) What is the expected pay-off for this game?

Answer: the pay-off is simply X_1+X_2. The expected pay-off is thus

$$\begin{aligned}\mathrm{E}[X_1+X_2]&=\sum_{x_1=0}^{1}\sum_{x_2=1}^{0}(x_1+x_2)f(x_1,x_2)\\&=0\cdot\frac{3}{8}+1\cdot\frac{2}{8}+1\cdot\frac{2}{8}+2\cdot\frac{1}{8}\\&=0.75.\end{aligned}$$

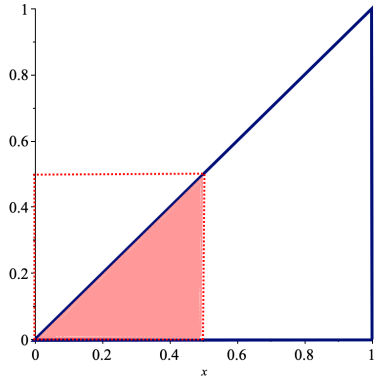
- Let X and Y have joint p.d.f.

$$f(x,y) = 2, \quad 0 \le y \le x \le 1.$$

a) What is the support of $f(x,y)$?

Answer: the support is the set $S = \{(x,y) : 0 \le y \le x \le 1\}$, a triangle in the xy plane bounded by the $x-$axis, the line $y = 1$, and the line $y = x$.

The support is the blue triangle shown below.



b) What is $P(0 \le X \le 0.5, 0 \le Y \le 0.5)$?

Answer: we need to evaluate the integral over the shaded area:

$$\begin{aligned} P(0 \le X \le 0.5, 0 \le Y \le 0.5) &= P(0 \le X \le 0.5, 0 \le Y \le X) \\ &= \int_0^{0.5} \int_0^x 2\,dy\,dx \\ &= \int_0^{0.5} [2y]_{y=0}^{y=x}\,dx \\ &= \int_0^{0.5} 2x\,dx = 1/4. \end{aligned}$$

c) What are the marginal probabilities $P(X = x)$ and $P(Y = y)$?

Answer: for $0 \le x \le 1$, we get

$$\begin{aligned} P(X = x) &= \int_{-\infty}^{\infty} f(x,y)\,dy \\ &= \int_{y=0}^{y=x} 2\,dy = [2y]_{y=0}^{y=x} = 2x, \end{aligned}$$

and for $0 \le y \le 1$,

$$\begin{aligned} P(Y = y) &= \int_{-\infty}^{\infty} f(x,y)\,dx = \int_{x=y}^{x=1} 2\,dx \\ &= [2x]_{x=y}^{x=1} = 2 - 2y. \end{aligned}$$

d) Compute $\mathrm{E}[X]$, $\mathrm{E}[Y]$, $\mathrm{E}[X^2]$, $\mathrm{E}[Y^2]$, and $\mathrm{E}[XY]$.

Answer: we have

$$\begin{aligned} \mathrm{E}[X] &= \iint_S x f(x,y)\,dA = \int_0^1 \int_0^x 2x\,dy\,dx \\ &= \int_0^1 [2xy]_{y=0}^{y=x}\,dx = \int_0^1 2x^2\,dx \\ &= \left[\frac{2}{3}x^3\right]_0^1 = \frac{2}{3}; \\ \mathrm{E}[Y] &= \iint_S y f(x,y)\,dA = \int_0^1 \int_y^1 2y\,dx\,dy \\ &= \int_0^1 [2xy]_{x=y}^{x=1}\,dy = \int_0^1 (2y - 2y^2)\,dy \\ &= \left[y^2 - \frac{2}{3}y^3\right]_0^1 = \frac{1}{3}; \\ \mathrm{E}[X^2] &= \iint_S x^2 f(x,y)\,dA = \int_0^1 \int_0^x 2x^2\,dy\,dx \\ &= \int_0^1 \left[2x^2 y\right]_{y=0}^{y=x}\,dx = \int_0^1 2x^3\,dx \\ &= \left[\frac{1}{2}x^4\right]_0^1 = \frac{1}{2}; \\ \mathrm{E}[Y^2] &= \iint_S y^2 f(x,y)\,dA = \int_0^1 \int_y^1 2y^2\,dx\,dy \\ &= \int_0^1 \left[2xy^2\right]_{x=y}^{x=1}\,dy = \int_0^1 (2y - 2y^3)\,dy \\ &= \left[\frac{2}{3}y^3 - \frac{1}{2}y^4\right]_0^1 = \frac{1}{6}; \\ \mathrm{E}[XY] &= \iint_S xyf(x,y) = \int_0^1 \int_0^x 2xy\,dy\,dx \\ &= \int_0^2 \left[xy^2\right]_{y=0}^{y=x} = \int_0^1 x^2\,dx \\ &= \left[\frac{x^4}{4}\right]_0^1 = \frac{1}{4}. \end{aligned}$$

e) Are X and Y independent?

Answer: they are not independent as the support of the joint p.d.f. is not rectangular.

The **covariance** of two random variables X and Y can give some indication of how they depend on one another:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]. \end{aligned}$$

When $X = Y$, the covariance reduces to the variance.[17]

Example: in the last example, $\mathrm{Var}[X] = \frac{1}{2} - (\frac{2}{3})^2 = \frac{1}{18}$, $\mathrm{Var}[X] = \frac{1}{6} - (\frac{1}{3})^2 = \frac{1}{18}$, and $\mathrm{Cov}(X,Y) = \frac{1}{4} - \frac{2}{3} \cdot \frac{1}{3} = \frac{1}{36}$.

[17] Note that the covariance could be negative, unlike the variance.

5. Descriptive Statistics

In a sense, the underlying reason for statistical analysis is to reach an **understanding of the data**.

5.1 Data Descriptions

Studies and experiments give rise to **statistical units**. These units are typically described with **variables** (and measurements), which are either **qualitative** (categorical) or **quantitative** (numerical).

Categorical variables take values (**levels**) from a finite set of pre-determined **categories** (or classes); numerical variables from a (potentially infinite) set of **quantities**.

Examples:

1. Age is a numerical variable, measured in years, although is is often reported to the nearest year integer, or in an age range of years, in which case it is an **ordinal** variable (mixture of qualitative or quantitative).
2. Typical numerical variables include distance in m, volume in cm^3, etc.
3. Disease diagnosis is a categorical variable with (at least) 2 categories (positive/negative).
4. Compliance with a standard is a categorical variable: there could be 2 levels (compliant/non-compliant) or more (compliance, minor non-compliance issues, major non-compliance issues).
5. Count variables are numerical variables.

Numerical Summaries In a first pass, a variable can be described along (at least) 2 dimensions: its **centrality** and its **spread** (the **skew** and the **kurtosis** are sometimes also used):

- **centrality** measures include the **median**, the **mean**, and, less frequently, the **mode**;
- spread (or **dispersion**) measures include the **standard deviation** (sd), the **quartiles**, the **inter-quartile range** (IQR), and, less frequently, the **range**.

The median, range, and quartiles are all easily calculated from an **ordered** list of the data.

Sample Median The **median** $\text{med}(x_1,\dots,x_n)$ of a sample of size n is a numerical value which splits the ordered data into 2 equal subsets: half the observations **below** below the median, and half **above** it:

- if n is **odd**, then the **position** of the median (or its **rank**) is $(n+1)/2$ – the median observation is the $\frac{n+1}{2}^{\text{th}}$ ordered observation;
- if n is **even**, then the median is the average of the $\frac{n}{2}^{\text{th}}$ and the $(\frac{n}{2}+1)^{\text{th}}$ ordered observations.

The procedure is simple: order the data, and follow the even/odd rules **to the letter**.

Examples

1. $\text{med}(4,6,1,3,7) = \text{med}(1,3,4,6,7) = x_{(5+1)/2} = x_3 = 4$. There are 2 observations below 4 $(1,3)$, and 2 observations above 4 $(6,7)$.
2. $\text{med}(1,3,4,6,7,23) = \frac{x_{6/2}+x_{6/2+1}}{2} = \frac{x_3+x_4}{2} = \frac{4+6}{2} = 5$. There are 3 observations below 5 $(1,3,4)$, and 3 observations above 4 $(6,7,23)$.
3. $\text{med}(1,3,3,6,7) = x_{(5+1)/2} = x_3 = 3$. There seems to be only 1 observation below 3 (1), but 2 observations above 3 $(6,7)$.

   This is not quite the correct interpretation of the median: **above** and **below** in the definition should be interpreted as **after** and **before**, respectively. In this example, there are 2 observations $(x_1 = 1, x_2 = 3)$ before the median $(x_3 = 3)$, and 2 after the median $(x_4 = 6, x_5 = 7)$.

Sample Mean The **mean** of a sample is simply the arithmetic average of its observations. For observations $x_1,\dots,x_n$, the sample mean is

$$\text{AM}(x_1,\dots,x_n) = \overline{x} = \frac{x_1+\cdots+x_n}{n} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

Other means exist, such as the **harmonic** mean and the **geometric** mean:

$$\text{HM}(x_1,\dots,x_n) = \frac{n}{\frac{1}{x_1}+\cdots+\frac{1}{x_n}}$$
$$\text{GM}(x_1,\dots,x_n) = \sqrt[n]{x_1\cdots x_n}.$$

Examples:

1. $\text{AM}(4,6,1,3,7) = \frac{4+6+1+3+7}{5} = \frac{21}{5} = 4.2 \approx 4 = \text{med}(4,6,1,3,7)$.
2. $\text{AM}(1,3,4,6,7,23) = \frac{1+3+4+6+7+23}{6} = \frac{44}{6} \approx 7.3$, which is not nearly as close to $\text{med}(1,3,4,6,7,23) = 5$.
3. $\text{HM}(4,6,1,3,7) = \frac{5}{\frac{1}{4}+\frac{1}{6}+\frac{1}{1}+\frac{1}{3}+\frac{1}{7}} = \frac{5}{53/28} = \frac{140}{53} \approx 2.64$.
4. $\text{GM}(4,6,1,3,7) = \sqrt[5]{4\cdot 6\cdot 1\cdot 3\cdot 7} \approx \sqrt[5]{(504)} \approx 3.47$.

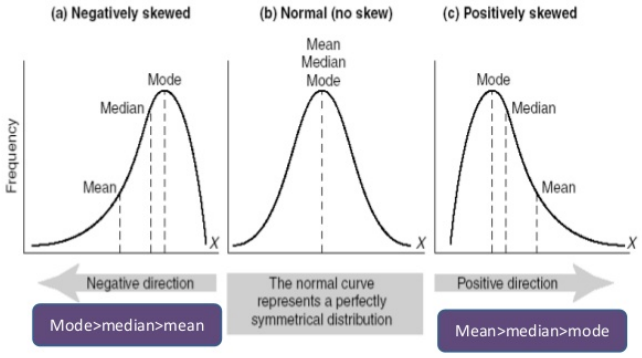
It can be shown that if $x = (x_1,\dots,x_n)$ and $x_i > 0$ for all i, then

$$\min(x) \le \text{HM}(x) \le \text{GM}(x) \le \text{AM}(x) \le \max(x).$$

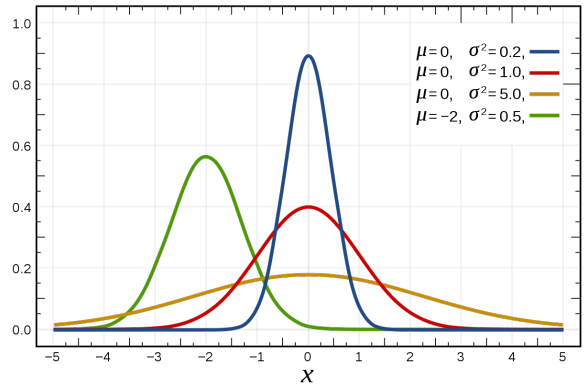
There is no need to decide on a single centrality measure when reporting on the data; in practice, we may use as many of them as we want to. But there are situations where the mean (or the median) could prove to be a better choice.

On the one hand, the use of the mean is **theoretically supported** by the Central Limit Theorem (see Section 6.2), and when the data distribution is roughly **symmetric**, then the median and the mean will be near one another.

If the data distribution is **skewed** then the mean is pulled toward the long tail and as a result gives a distorted view of the centre. Consequently, medians are generally used for house prices, incomes, etc., as the median is **robust** against outliers and incorrect readings (whereas the mean is not).



Standard Deviation While the mean, the median, and the mode provide an idea as to where some of the distribution's "mass" is located, the **standard deviation** provides some notion of its spread. The higher the standard deviation, the further away from the mean the variable values are likely to fall (see below). We will have more to say on this topic.



Quartiles Another way to provide information about the spread of the data is *via* **centiles**, **deciles**, and/or **quartiles**.

The **lower quartile** $Q_1(x_1,\ldots,x_n)$ of a sample of size n, or Q_1, is a numerical value which splits the ordered data into 2 unequal subsets: 25% of the observations fall below Q_1 **and** 75% of the observations fall above Q_1.

Similarly, the **upper quartile** Q_3 splits the ordered data into 75% of the observations below Q_3, **and** 25% of the observations above Q_3.

The median can be interpreted as the **middle quartile** Q_2, of the sample, the minimum as Q_0, and the maximum as Q_4: the vector (Q_0,Q_1,Q_2,Q_3,Q_4) is the **5-pt summary** of the data.

Centiles p_i, $i=0,\ldots,100$ and deciles d_j, $j=0,\ldots,10$ run through different splitting percentages

$$p_{25}=Q_1, p_{75}=Q_3, d_5=Q_2, \text{ etc.}$$

Sort the sample observations $\{x_1,x_2,\ldots,x_n\}$ in an **increasing order** as

$$y_1\le y_2\le\ldots\le y_n.$$

The smallest y_1 has **rank** 1 and the largest y_n has **rank** n.

Any value that falls between the observations of ranks:

- $\lfloor\frac{n}{4}\rfloor$ and $\lfloor\frac{n}{4}\rfloor+1$ is a **lower quartile** Q_1;
- $\lfloor\frac{3n}{4}\rfloor$ and $\lfloor\frac{3n}{4}\rfloor+1$ is an **upper quartile** Q_3;
- $\lfloor\frac{in}{100}\rfloor$ and $\lfloor\frac{in}{100}\rfloor+1$ is a **centile** p_i, for $i=1,\ldots,99$;
- $\lfloor\frac{jn}{10}\rfloor$ and $\lfloor\frac{jn}{10}\rfloor+1$ is a **decile** d_j, for $j=1,\ldots,9$.

In practice, we compute the m**—quantile of order** k for the data, where $k=1,\ldots,m-1$ by averaging the observations of rank

$$\left\lfloor\frac{kn}{m}\right\rfloor \quad\text{and}\quad \left\lfloor\frac{kn}{m}\right\rfloor+1.$$

Examples

$$\begin{aligned}Q_1(1,3,4,6,7)&=\frac{1}{2}\left(y_{\lfloor 5/4\rfloor}+y_{\lfloor 5/4\rfloor+1}\right)\\&=\frac{1}{2}\left(y_1+y_2\right)\\&=\frac{1}{2}(1+3)=2;\\ d_7(1,3,4,6,7,23)&=\frac{1}{2}\left(y_{\lfloor 7(6)/10\rfloor}+y_{\lfloor 7(6)/10\rfloor+1}\right)\\&=\frac{1}{2}\left(y_4+y_5\right)\\&=\frac{1}{2}(6+7)=13/2.\end{aligned}$$

Dispersion Measures Some of the dispersion measures are fairly simple to compute: the **sample range** is

$$\text{range}(x_1,\ldots,x_n)=\max\{x_i\}-\min\{x_i\};$$

the **inter-quartile range** is $\text{IQR}=Q_3-Q_1$.

The **sample standard deviation** s and **sample variance** s^2 are estimates of the underlying distribution's σ and σ^2. For observations $x_1,\ldots,x_n$,

$$s^2=\frac{1}{n-1}\sum_{i=1}^n(x_i-\overline{x})^2=\frac{1}{n-1}\left(\sum_{i=1}^n x_i^2-\frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2\right);$$

it differs from the (population) standard deviation and the (population) variance in the denominator: $n-1$ is used instead of n.[18]

Example: the sample variance of $\{1,3,4,6,7\}$ is

$$\frac{1}{5-1}\left(\sum_{i=1}^5 x_i^2-\frac{1}{5}\left(\sum_{i=1}^5 x_i\right)^2\right)=\frac{1}{4}\left(111-\frac{1}{5}(21)^2\right)=5.7.$$

[18] In statistical parlance, we say that 1 degree of freedom is lost when we use the sample to estimate the sample mean.

Outliers An **outlier** is an observation that lies outside the overall pattern in a distribution.[19]

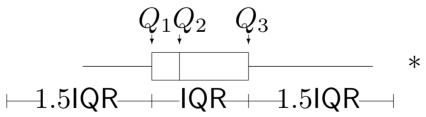
Let x be an observation in the sample;[20] it is a

- **suspected outlier** if
$$x < Q_1 - 1.5\,\text{IQR} \quad \text{or} \quad x > Q_3 + 1.5\,\text{IQR},$$
- **definite outlier** if
$$x < Q_1 - 3\,\text{IQR} \quad \text{or} \quad x > Q_3 + 3\,\text{IQR}.$$

5.2 Visual Summaries

The **boxplot** (also known as the box-and-whisker plot) is a quick and easy way to present a graphical summary of a univariate distribution:

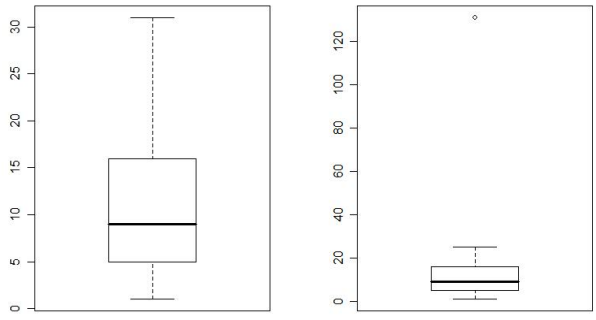
1. draw a box along the observation axis, with endpoints at the lower and upper quartiles Q_1 and Q_3, and with a "belt" at the median Q_2;
2. draw a line extending from Q_1 to the smallest value closer than 1.5IQR to the left of Q_1;
3. draw a line extending from Q_3 to the largest value closer than 1.5IQR to the right of Q_3;
4. any suspected outlier is plotted separately (as below):



Skewness For **symmetric** distributions, the median and mean are equal, and the Q_1 and Q_3 are equidistant from Q_2:

- if $Q_3 - Q_2 > Q_2 - Q_1$ then the data distribution is **skewed to the right**;
- if $Q_3 - Q_2 < Q_2 - Q_1$ then the data distribution is **skewed to left**.

In the boxplots below, the data is skewed to the right.

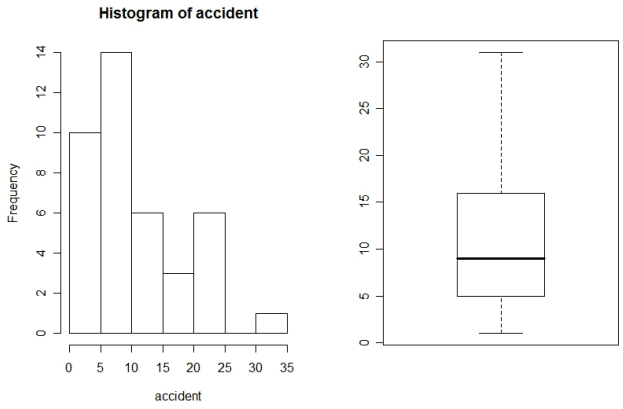


Histograms Visual information can about the distribution of the sample can also be provided *via* **histograms**.

A histogram for the sample $\{x_1, \ldots, x_n\}$ is built according to the following specifications:

- the **range** of the histogram is $r = \max\{x_i\} - \min\{x_i\}$;
- the **number of bins** should approach $k = \sqrt{n}$, where n is the sample size;
- the **bin width** should approach r/k, and
- the **frequency of observations** in each bin should be represented by the **bin height**.

Shapes of Datasets Boxplots and histograms provide an easy visual impression of the **shape of the data set**, which can eventually suggest a mathematical model for the situation of interest: another way to define skewness is to say that data is said to be **skewed to the right** if the corresponding boxplot or histogram is stretched to the right.



5.3 Coefficient of Correlation

Consider the following data, consisting of $n = 20$ paired measurements (x_i, y_i) of hydrocarbon levels (x) and pure oxygen levels (y) in fuels:

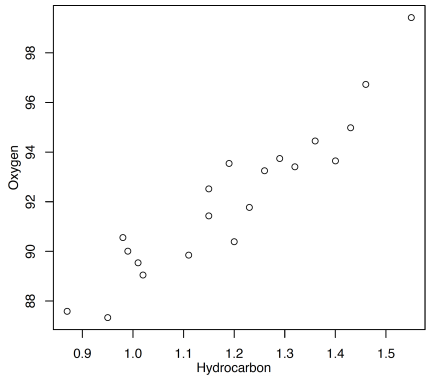
```
x:  0.99  1.02  1.15  1.29  1.46
y: 90.01 89.05 91.43 93.74 96.73

x:  1.36  0.87  1.23  1.55  1.40
y: 94.45 87.59 91.77 99.42 93.65

x:  1.19  1.15  0.98  1.01  1.11
y: 93.54 92.52 90.56 89.54 89.85

x:  1.20  1.26  1.32  1.43  0.95
y: 90.39 93.25 93.41 94.98 87.33
```

Assume that we are interested in measuring the **strength of association** between x and y.

We can use a graphical display to provide an initial description of the relationship: it appears that the observations lie around a **hidden line**.

[19] Outlier analysis (and anomaly detection) is its own discipline – an overview is provided here [1].

[20] In theory, this definition only applies to **normally distributed** data, but it is often used as a first pass during outlier analysis even when the data is not normally distributed.



For paired data (x_i, y_i), $i = 1, \ldots, n$, the **sample correlation coefficient** of x and y is

$$\rho_{XY} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}.$$

The coefficient ρ_{XY} is defined only if $S_{xx} \neq 0$ and $S_{yy} \neq 0$, i.e. neither x_i nor y_i are constant.

The variables x and y are **uncorrelated** if $\rho_{XY} = 0$ (or very small, in practice), and **correlated** if $\rho_{XY} \neq 0$ (or $|\rho_{XY}|$ is "large", in practice).

Example: for the data on the previous page, we have

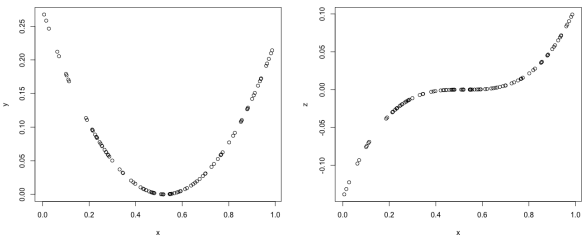
$$S_{xy} \approx 10.18, \; S_{xx} \approx 0.68, \; S_{yy} \approx 173.38,$$

so that

$$\rho_{XY} \approx \frac{10.18}{\sqrt{0.68 \cdot 173.38}} \approx 0.94.$$

Properties

- ρ_{XY} is unaffected by changes of scale or origin. Adding constants to x does not change $x - \overline{x}$ and multiplying x and y by constants changes both the numerator and denominator equally;
- ρ_{XY} is symmetric in x and y (i.e. $\rho_{XY} = \rho_{YX}$) and $-1 \leq \rho_{XY} \leq 1$; if $\rho_{XY} = \pm 1$, then the observations (x_i, y_i) all lie on a straight line with a positive (negative) slope;
- the sign of ρ_{XY} reflects the trend of the points;
- a high correlation coefficient value $|\rho_{XY}|$ does not necessarily imply a **causal relationship** between the two variables;
- note that x and y can have a very strong **non-linear** relationship without ρ_{XY} reflecting it (−0.12 on the left, 0.93 on the right).



6. Central Limit Theorem and Sampling Distributions

In this section, we introduce one of the fundamental results of probability theory and statistical analysis.

6.1 Sampling Distributions

A **population** is a set of similar items which of interest in relation to some questions or experiments.

In some situations, it is impossible to observe the entire set of observations that make up a population – perhaps the entire population is too large to query, or some units are out-of-reach.

In these cases, we can only hope to infer the behaviour of the entire population by considering a **sample** (subset) of the population.

Suppose that $X_1, \ldots, X_n$ are n **independent** random variables, each having the same c.d.f. F, i.e. they are **identically distributed**. Then, $\{X_1, \ldots, X_n\}$ is a **random sample** of size n from the population, with c.d.f. F.

Any function of such a random sample is called a **statistic** of the sample; the probability distribution of a statistic is called a **sampling distribution**.

Recall the linear properties of the expectation and the variance: if X is a random variable and $a, b \in \mathbb{R}$, then

$$\begin{aligned} \mathrm{E}[a + bX] &= a + b\mathrm{E}[X], \\ \mathrm{Var}[a + bX] &= b^2 \mathrm{Var}[X], \\ \mathrm{SD}[a + bX] &= |b| \, \mathrm{SD}[X]. \end{aligned}$$

Sum of Independent Random Variables For any random variables X and Y, we have

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

In general,

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + 2\mathrm{Cov}(X, Y) + \mathrm{Var}[Y];$$

if **in addition** X and Y are **independent**, then

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

More generally, if $X_1, X_2, \ldots, X_n$ are **independent**, then

$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{E}[X_i] \quad \text{and} \quad \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i].$$

Independent and Identically Distributed Random Variables A special case of the above occurs when all of $X_1, \ldots, X_n$ have **exactly the same distribution**. In that case we say they are **independent and identically distributed**, which is traditionally abbreviated to "**iid**".

If $X_1, \ldots, X_n$ are iid, and

$$\mathrm{E}[X_i] = \mu \quad \text{and} \quad \mathrm{Var}[X_i] = \sigma^2 \quad \text{for } i = 1, \ldots, n,$$

then

$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = n\mu \quad \text{and} \quad \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = n\sigma^2 .$$

Examples

- A random sample of size 100 is taken from a population with mean 50 and variance 0.25. Find the expected value and variance of the **sample total**.

  Answer: this problem translates to "if $X_1, \ldots, X_{100}$ are iid with $\mathrm{E}[X_i] = \mu = 50$ and $\mathrm{Var}[X] = \sigma^2 = 0.25$ for $i = 1, \ldots, 100$, find $\mathrm{E}[\tau]$ and $\mathrm{Var}[\tau]$ for

  $$\tau = \sum_{i=1}^{n} X_i ."$$

  According to the iid formulas,

  $$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = 100\mu = 5000$$
  $$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = 100\sigma^2 = 25 .$$

- The mean value of potting mix bags weights is 5 kg, with standard deviation 0.2. If a shop assistant carries 4 bags (selected independently from stock) then what is the expected value and standard deviation of the total weight carried?

  Answer: there is an implicit "population" of bag weights. Let X_1, X_2, X_3, X_4 be iid with $\mathrm{E}[X_i] = \mu = 5$, $\mathrm{SD}[X_i] = \sigma = 0.2$ and $\mathrm{Var}[X_i] = \sigma^2 = 0.2^2 = 0.04$ for $i = 1, 2, 3, 4$. Let $\tau = X_1 + X_2 + X_3 + X_4$.

  According to the iid formulas,

  $$\mathrm{E}[\tau] = n\mu = 4 \cdot 5 = 20$$
  $$\mathrm{Var}[\tau] = n\sigma^2 = 4 \cdot 0.04 = 0.16.$$

  Thus, $\mathrm{SD}[\tau] = \sqrt{0.16} = 0.4$.

Sample Mean (Reprise) The **sample mean** is a typical statistic of interest:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i .$$

If $X_1, \ldots, X_n$ are iid with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$ for all $i = 1, \ldots, n$, then

$$\mathrm{E}\left[\overline{X}\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}(n\mu) = \mu$$
$$\mathrm{Var}\left[\overline{X}\right] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n} .$$

Example: a set of scales returns the true weight of the object being weighed plus a random error with mean 0 and standard deviation 0.1 g. Find the standard deviation of the average of 9 such measurements of an object.

Answer: suppose the object has true weight μ. The "random error" indicates that each measurement $i = 1, \ldots, 9$ is written as $X_i = \mu + Z_i$ where $\mathrm{E}[Z_i] = 0$ and $\mathrm{SD}[Z_i] = 0.1$ and the Z_i's are iid.

The X_i's are iid with $\mathrm{E}[X_i] = \mu$ and $\mathrm{SD}[X_i] = \sigma = 0.1$. If we average $X_1, \ldots, X_n$ (with $n = 9$) to get $\overline{X}$, then

$$\mathrm{E}\left[\overline{X}\right] = \mu \text{ and } \mathrm{SD}\left[\overline{X}\right] = \frac{\sigma}{\sqrt{n}} = \frac{0.1}{\sqrt{9}} = \frac{1}{30} \approx 0.033 .$$

We do not need to know the **actual** distribution of the X_i; only μ and σ^2 are required to compute $\mathrm{E}[\overline{X}]$ and $\mathrm{Var}[\overline{X}]$.

Sum of Independent Normal Random Variables Another interesting case occurs when we have **multiple independent normal** random variables on the same experiment.

Suppose $X_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ for $i = 1, \ldots, n$, and all the X_i are independent. We already know that

$$\begin{aligned}\mathrm{E}[\tau] &= \mathrm{E}[X_1 + \cdots + X_n] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n] \\ &= \mu_1 + \cdots + \mu_n ; \\ \mathrm{Var}[\tau] &= \mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] \\ &= \sigma_1^2 + \cdots + \sigma_n^2 .\end{aligned}$$

It turns out that, under these hypotheses, τ is **also normally distributed**, i.e.

$$\tau = \sum_{i=1}^{n} X_i \sim \mathcal{N}(\mathrm{E}[\tau], \mathrm{Var}[\tau]) = \mathcal{N}\left(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2\right).$$

Thus, if $\{X_1, \ldots, X_n\}$ is a random sample from a normal population **with mean μ and variance σ^2**, then $\sum_{i=1}^{n} X_i$ and $\overline{X}$ are also normal, which, combined with the above work, means that

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right) \quad \text{and} \quad \overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) .$$

Example: suppose that the population of students' weights is normal with mean 75 kg and standard deviation 5 kg. If 16 students are picked at random, what is the distribution of the (random) total weight τ? What is the probability that the total weight exceeds 1250 kg?

Answer: If $X_1, \ldots, X_{16}$ are iid as $\mathcal{N}(75, 25)$, then the sum $\tau = X_1 + \cdots + X_{16}$ is also normally distributed with

$$\tau = \sum_{i=1}^{16} X_i \sim \mathcal{N}(16 \cdot 75, 16 \cdot 25) = \mathcal{N}(1200, 400), \quad \text{and}$$

$$Z = \frac{\tau - 1200}{\sqrt{400}} \sim \mathcal{N}(0, 1).$$

Thus,

$$P(\tau > 1250) = P\left(\frac{\tau - 1200}{\sqrt{400}} > \frac{1250-1200}{20}\right)$$
$$= P(Z > 2.5) = 1 - P(Z \leq 2.5)$$
$$\approx 1 - 0.9938 = 0.0062.$$

6.2 Central Limit Theorem

Suppose that a professor has been teaching a course for the last 20 years. For every cohort during that period, the mid-term exam grades of all the students have been recorded.

Let $X_{i,j}$ be the grade of student i in year j. Looking back on the class lists, they find that

$$\mathrm{E}[X_{i,j}] = 56 \quad \text{and} \quad \mathrm{SD}[X_{i,j}] = 11.$$

This year, there are 49 students in the class. What should the professor expect for the class mid-term exam average?

Of course, the professor cannot predict any of the student grades or the class average with absolute certainty, but they could try the following approach:

1. simulate the results of the class of 49 students by generating sample grades $X_{1,1}, \ldots, X_{1,49}$ from a **normal** distribution $\mathcal{N}(65, 15^2)$;
2. compute the sample mean for the sample and record it as $\overline{X}_1$;
3. repeat steps 1-2 m times and compute the standard deviation of the sample means $\overline{X}_1, \ldots, \overline{X}_m$;
4. plot the histogram of the sample means $\overline{X}_1, \ldots, \overline{X}_m$.

What do you think is going to happen?

Central Limit Theorem: If $\overline{X}$ is the mean of a random sample of size n taken from an **unknown** population with mean μ and finite variance σ^2, then

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1),$$

as $n \to \infty$.

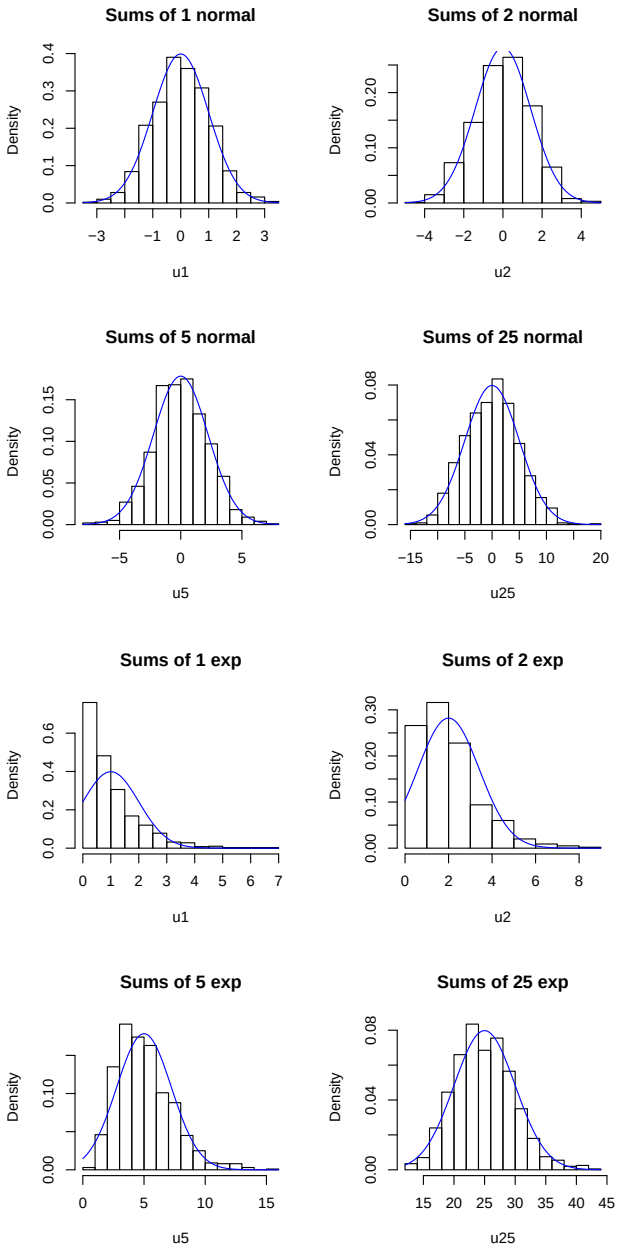
More precisely, this is a **limiting** result. If we view the **standardization**

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

as functions of n, we have, for each z,

$$\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z) \text{ and}$$
$$P(Z_n \leq z) \approx \Phi(z), \text{ if } n \text{ is large enough,}$$

whether the original X_i's are normal or not.



Examples

- The examination scores in an university course have mean 56 and standard deviation 11. In a class of 49 students, what is the probability that the average mark is below 50? What is the probability that the average mark lies between 50 and 60?

  Answer: let the marks be $X_1, \ldots, X_{49}$ and assume the performances are independent. According to the central limit theorem,

  $$\overline{X} = (X_1 + X_2 + \cdots + X_{49})/49,$$

  with $\mathrm{E}[\overline{X}] = 56$ and $\mathrm{Var}[\overline{X}] = 11^2/49$.

We thus have

$$P(\overline{X} < 50) \approx P\left(Z < \frac{50-56}{11/7}\right) = P(Z < -3.82) = 0.0001$$

and

$$P(50 < \overline{X} < 60) \approx P\left(\frac{50-56}{11/7} < Z < \frac{60-56}{11/7}\right) = P(-3.82 < Z < 2.55) = \Phi(2.55) - \Phi(-3.82) = 0.9945.$$

Note that this says nothing about whether the scores are normally distributed or not, only that the average scores follow an approximate normal distribution.[21]

- Systolic blood pressure readings for pre-menopausal, non-pregnant women aged $35-40$ have mean 122.6 standard deviation 11 mm Hg. An independent sample of 25 women is drawn from this target population and their blood pressure is recorded.

  What is the probability that the average blood pressure is greater than 125 mm Hg? How would the answer change if the sample size increases to 40?

  Answer: according to the CLT, $\overline{X} \sim \mathcal{N}(122.6, 121/25)$, approximately. Thus

$$P(\overline{X} > 125) \approx P\left(Z > \frac{125-122.6}{11/\sqrt{25}}\right) = P(Z > 1.09) = 1 - \Phi(1.09) = 0.14.$$

  However, if the sample size is 40, then

$$P(\overline{X} > 125) \approx P\left(Z > \frac{125-122.6}{11/\sqrt{40}}\right) = 0.08.$$

  Increasing the sample size reduces the probability that the average is far from the expectation of each original measurement.

- Suppose that we select a random sample $X_1, \ldots, X_{100}$ from a population with mean 5 and variance 0.01.

  What is the probability that the difference between the sample mean of the random sample and the mean of the population exceeds 0.027?

  Answer: according to the CLT, we know that, approximately, $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ has standard normal distribution.

The desired probability is thus

$$\begin{aligned} P &= P(|\overline{X} - \mu| \geq 0.027) \\ &= P(\overline{X} - \mu \geq 0.027 \text{ or } \mu - \overline{X} \geq 0.027) \\ &= P\left(\frac{\overline{X}-5}{0.1/\sqrt{100}} \geq \frac{0.027}{0.1/\sqrt{100}}\right) + P\left(\frac{\overline{X}-5}{0.1/\sqrt{100}} \leq \frac{-0.027}{0.1/\sqrt{100}}\right) \\ &\approx P(Z \geq 2.7) + P(Z \leq -2.7) \\ &= 2P(Z \geq 2.7) \approx 2(0.0035) = 0.007. \end{aligned}$$

6.3 Sampling Distributions (Reprise)

We now revisit sampling distributions.

Difference Between Two Means Statisticians are often interested in the difference between various populations; a result akin to the central limit theorem provides guidance in that area.

Theorem: Let $\{X_1, \ldots, X_n\}$ be a random sample from a population with mean μ_1 and variance σ_1^2, and $\{Y_1, \ldots, Y_m\}$ be another random sample, independent of X, from a population with mean μ_2 and variance σ_2^2.

If $\overline{X}$ and $\overline{Y}$ are the respective sample means, then

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

has standard normal distribution $\mathcal{N}(0,1)$ as $n, m \to \infty$.[22]

Example: two different machines are used to fill cereal boxes on an assembly line. The critical measurement influenced by these machines is the weight of the product in the boxes.

The variances of these weights is identical, $\sigma^2 = 1$. Each machine produces a sample of 36 boxes, and the weights are recorded. What is the probability that the difference between the respective averages is less than 0.2, assuming that the true means are identical?

Answer: we have $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2 = 1$, and $n = m = 36$. The desired probability is

$$\begin{aligned} P\left(|\overline{X} - \overline{Y}| < 0.2\right) &= P\left(-0.2 < \overline{X} - \overline{Y} < 0.2\right) \\ &= P\left(\frac{-0.2-0}{\sqrt{1/36+1/36}} < \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{1/36+1/36}} < \frac{0.2-0}{\sqrt{1/36+1/36}}\right) \\ &= P(-0.8485 < Z < 0.8485) \\ &\approx \Phi(0.8485) - \Phi(-0.8485) \approx 0.6. \end{aligned}$$

[21] If the scores did arise from a normal distribution, the $\approx$ would be replaced by a $=$, as per Section 6.1.

[22] Like the CLT, this is a **limiting** result.

Sample Variance S^2 When the underlying variance is unknown (which is usually the case in practice), it must be approximated by the sample variance.

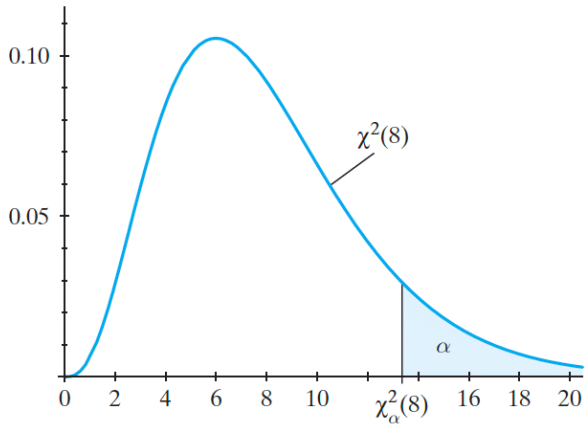
Theorem: Let $\{X_1,\dots,X_n\}$ be a random sample taken from a normal population with mean σ^2, and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

be the sample variance. The statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2}$$

follows a **chi-squared distribution with $\nu = n-1$ degrees of freedom** (d.f.), where $\chi^2(\nu) = \Gamma(1/2,\nu)$.



Notation: for $0<\alpha<1$ and $\nu\in\mathbb{N}^*$, $\chi^2_\alpha(\nu)$ is the **critical value** for which

$$P(\chi^2 > \chi^2_\alpha(\nu)) = \alpha,$$

where $\chi^2\sim\chi^2(\nu)$ follows a chi-squared distribution with ν degrees of freedom.

The values of $\chi^2_\alpha(\nu)$ can be found in various textbook tables, or by using `R` or specialized online calculators.

Example: for instance, when $\nu = 8$ and $\alpha = 0.95$, we have

$$\chi^2_{0.95}(8) = 2.732,$$ [23]

therefore $P(\chi^2 > 2.732) = 0.95$, where $\chi^2\sim\chi^2(8)$, i.e. χ^2 has a chi-squared distribution with $\nu = 8$ degrees of freedom.

In other words, 95% of the area under the curve of the probability density function of $\chi^2(8)$ is found to the right of 2.732.

[23] `qchisq(0.95, df=8,lower.tail = FALSE)`

Sample Mean With Unknown Population Variance Suppose that $Z\sim\mathcal{N}(0,1)$ and $V\sim\chi^2(\nu)$. If Z and V are independent, then the distribution of the random variable

$$T = \frac{Z}{\sqrt{V/\nu}}$$

is a **Student t−distribution with ν degrees of freedom**, which we denote by $T\sim t(\nu)$.[24]

Theorem: let $X_1,\dots,X_n$ be independent normal random variables with mean μ and standard deviation σ. Let $\overline{X}$ and S^2 be the sample mean and sample variance, respectively. Then the random variable

$$T = \frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1),$$

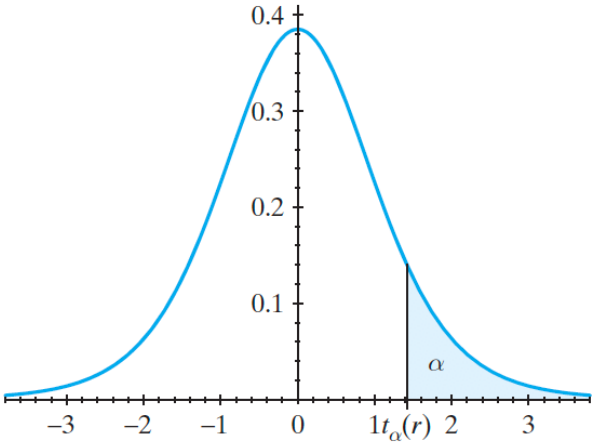
follows a **Student t−distribution with $\nu = n-1$ degrees of freedom**.

Using the same notation as with the chi-squared distribution, let $t_\alpha(\nu)$ represent the **critical t-value** above which we find an area under the p.d.f. of $t(\nu)$ equal to α, i.e.

$$P(T > t_\alpha(\nu)) = \alpha,$$

where $T\sim t(\nu)$.

For all ν, the Student t-distribution is a symmetric distribution around zero, so we have $t_{1-\alpha}(\nu) = -t_\alpha(\nu)$. The critical values can be found in tables, or by using the `R` function `qt()`.



If $T\sim t(\nu)$, then for any $0<\alpha<1$, we have

$$\begin{aligned}
P\left(|T| < t_{\alpha/2}(\nu)\right) &= P\left(-t_{\alpha/2}(\nu) < T < t_{\alpha/2}(\nu)\right)\\
&= P\left(T < t_{\alpha/2}(\nu)\right) - P\left(T < -t_{\alpha/2}(\nu)\right)\\
&= 1 - P\left(T > t_{\alpha/2}(\nu)\right) - \left(1 - P\left(T > -t_{\alpha/2}(\nu)\right)\right)\\
&= 1 - P\left(T > t_{\alpha/2}(\nu)\right) - \left(1 - P\left(T > t_{1-\alpha/2}(\nu)\right)\right)\\
&= 1 - \alpha/2 - (1-(1-\alpha/2)) = 1-\alpha.
\end{aligned}$$

[24] The probability density function of $t(\nu)$ is

$$f(x) = \frac{\Gamma(\nu/2+1/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)(1+x^2/\nu)^{\nu/2+1/2}}.$$

Consequently,

$$P\left(-t_{\alpha/2}(n-1) < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1)\right) = 1-\alpha\,.$$

We can show that $t(\nu) \to \mathcal{N}(0,1)$ as $\nu \to \infty$; intuitively, this makes sense because the estimate S gets better at estimating σ when n increases.

Example: in R, we can see that when $T \sim t(8)$,

$$P(T > 2.306) = 0.025,$$ [25]

which implies $P(T < -2.306) = 0.025$, so $t_{0.025}(8) = 2.306$ and

$$\begin{aligned} P(|T| \leq 2.306) &= P(-2.306 \leq T \leq 2.306) \\ &= 1 - P(T < -2.306) - P(T > 2.306) \\ &= 1 - 2P(T < -2.306) = 0.95\,. \end{aligned}$$

The Student $t-$distribution will be useful when the time comes to compute confidence intervals and to do hypothesis testing (see Sections 7 and 8).

$F-$Distributions Let $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$. If U and V are independent, then the random variable

$$F = \frac{U/\nu_1}{V/\nu_2}$$

follows an **F-distribution with ν_1 and ν_2 degrees of freedom**, which we denote by $F \sim F(\nu_1, \nu_2)$.

The probability density function of $F(\nu_1, \nu_2)$ is

$$f(x) = \frac{\Gamma(\nu_1/2+\nu_2/2)(\nu_1/\nu_2)^{\nu_1/2}x^{\nu_1/2-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)(1+x\,\nu_1/\nu_2)^{\nu_1/2+\nu_2/2}}, \quad x \geq 0.$$

Theorem: If S_1^2 and S_2^2 are the sample variances of independent random samples of size n and m, respectively, taken from normal populations with variances σ_1^2 and σ_2^2, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

follows an **F-distribution with $\nu_1 = n-1$, $\nu_2 = m-1$ d.f.**

Notation: for $0 < \alpha < 1$ and $\nu_1, \nu_2 \in \mathbb{N}^*$, $f_\alpha(\nu_1, \nu_2)$ is the **critical value** for which $P(F > f_\alpha(\nu_1, \nu_2)) = \alpha$ where $F \sim F(\nu_1, \nu_2)$. Critical values can be found in tables, or by using the R function `qf()`.

It can be shown that

$$f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_2, \nu_1)};$$

for instance, $f_{0.95}(6,10) = \frac{1}{f_{0.05}(10,6)} = \frac{1}{4.06} = 0.246\,.$ [26]

These distributions play a role in linear regression and ANOVA models (see Section 9).

[25] `qt(0.025, df=8, lower.tail=FALSE)`
[26] `qf(0.95, df1=6, df2=10, lower.tail=FALSE)`

7. Point and Interval Estimation

Statistical inference (generalizing from a sample to the population) is one of the objectives of statistical analysis.

7.1 Statistical Inference

One of the goals of **statistical inference** is to draw conclusions about a **population** based on a random sample from the population.

Examples

- Can we assess the reliability of a product's manufacturing process by randomly selecting a sample of the final product and determining how many of them are compliant according to some quality assessment scheme?
- Can we determine who will win an election by polling a small sample of respondents?

Specifically, we seek to estimate an unknown **parameter** θ, say, using a single quantity called the **point estimate** $\hat{\theta}$.

This point estimate is obtained *viar* a **statistic**, which is simply a function of a random sample.

The probability distribution of the statistic is its **sampling distribution**; as an example, we have discussed the sampling distribution of the **sample mean** in the previous section. Describing such sampling distributions is a main area of research.

Example: consider a process that manufactures gear wheels (in some standard gauge). Let X be the random variable that records the weight of a randomly selected gear wheel. What is the population mean $\mu_X = \mathrm{E}[X]$?.

Answer: in the absence of the p.d.f. $f(x)$, we can estimate $\mu = X$ with the help of a random sample $X_1, \ldots, X_n$ of gear wheel weight measurements, *via* the sample mean statistic:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n},$$

which approximately follows $\mathcal{N}\left(\mu, \sigma^2/n\right)$, according to the central limit theorem.

Statistics Common examples of statistics include:

- the sample mean and the sample median;
- the sample variance and the sample standard deviation;
- sample quantiles (median, quartiles, quantiles);
- test statistics ($t-$statistics, χ^2-statistics, $f-$statistics, etc.);
- order statistics (sample maximum and minimum, sample range, etc.);
- sample moments and functions thereof (skewness, kurtosis, etc.);
- etc.

Estimator Variance and Standard Error In practice, the point estimator $\hat{\theta}$ varies depending on the choice of the sample $\{X_1,\ldots,X_n\}$.

The **standard error** of a statistic is the **standard deviation of its sampling distribution**.

For instance, if observations $X_1,\ldots,X_n$ come from a a population with **unknown mean** μ and **known variance** σ^2, then $\text{Var}(\overline{X}) = \sigma^2/n$ and the **standard error of** $\overline{X}$ is

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

If the variance of the original population is **unknown**, then it is estimated by the sample variance S^2 and the **estimated standard error of** $\overline{X}$ is

$$\hat{\sigma}_{\overline{X}} = \frac{S}{\sqrt{n}}, \quad \text{where} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Examples

1. A sample of 20 baseball player heights (in inches) is shown below.

   ```
   74,74,72,72,73,69,69,71,76,71,
   73,73,74,74,69,70,72,73,75,78.
   ```

   What is the standard error of the sample mean $\overline{X}$?

   Answer: the sampling mean of the heights is

   $$\overline{X} = \frac{X_1 + \cdots + X_{20}}{20} = 72.6$$

   and the sample variance S^2 is

   $$S^2 = \frac{1}{20-1}\sum_{i=1}^{20}(X_i - 72.6)^2 \approx 5.6211.$$

   The standard error of $\overline{X}$ is thus

   $$\hat{\sigma}_{\overline{X}} = \frac{S}{\sqrt{20}} \approx \sqrt{\frac{5.6211}{20}} \approx 0.5301.$$

2. Consider a sample $\{X_1,\ldots,X_{100}\}$ of independent observations selected from a normal population $\mathcal{N}(\mu,\sigma^2)$ where $\sigma = 50$ is known, but μ is not. What is the best estimate of μ? What is the sampling distribution of that estimate?

   Answer: the sample mean $\overline{X} = \frac{X_1+\cdots+X_{100}}{100}$ provides the best estimate of $\mu_X = \mu_{\overline{X}}$ and the standard error of $\overline{X}$ is $\sigma_{\overline{X}} = \frac{50}{\sqrt{100}} = 5$.

   Since the observations are sampled independently from a normal population with mean μ and standard deviation 50, $\overline{X} \sim \mathcal{N}(\mu, 5^2) = \mathcal{N}(\mu, 25)$, according to the CLT.

7.2 Confidence Interval for μ when σ is Known

In general, consider a sample $\{x_1,\ldots,x_n\}$ from a **normal population** with **known** variance σ^2 and **unknown** mean μ. The sample mean

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

is a **point estimate** of μ.[27]

Of course, this estimate is not exact, because $\overline{x}$ is an **observed value** of $\overline{X}$; it is unlikely that the observed value $\overline{x}$ should coincide with μ.

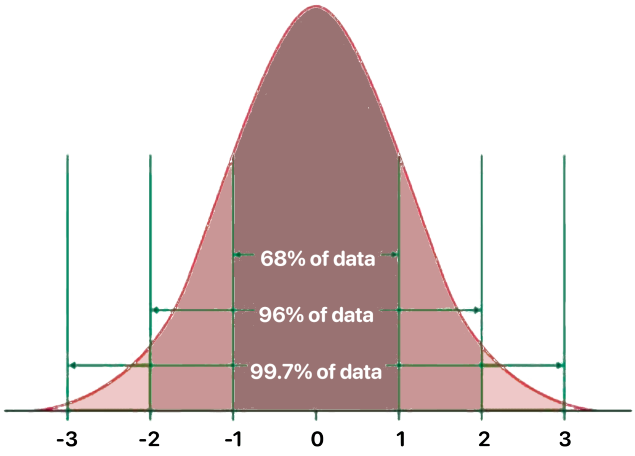
We know that $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, so that

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

The $68-96-99.7$ **Rule** For the standard normal distribution, it can be shown that:

$$P(|Z| < 1) \approx 0.683,\ P(|Z| < 2) \approx 0.955,\ P(|Z| < 3) \approx 0.997.$$

This says that about 68% of the observations from $\mathcal{N}(0,1)$ fall within one standard deviation ($\sigma = 1$) from the mean ($\mu = 0$), about 96% within two standard deviations, and about 99.7% within three.



In other words, whenever we observe a sample mean $\overline{X}$ (with sample size n) from a normal population with mean μ, we would expect the inequality

$$-k < Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < k$$

to hold approximately

$$g(k) = \begin{cases} 68.3\% \text{ of the time}, & \text{if } k = 1 \\ 95.5\% \text{ of the time}, & \text{if } k = 2 \\ 99.7\% \text{ of the time}, & \text{if } k = 3 \end{cases}$$

[27] In general, upper case letters are reserved for a general sample, and lower case letters for a specifically observed sample.

Confidence Intervals By re-arranging the terms, we can build a **symmetric** $g(k)$ **confidence interval** (C.I.) **for** μ:

$$\overline{X} - k\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + k\frac{\sigma}{\sqrt{n}} \implies \text{C.I.}(\mu; g(k)) \equiv \overline{X} \pm k\frac{\sigma}{\sqrt{n}}.$$

Examples

- Consider a sample $\{X_1, \ldots, X_{64}\}$ from a normal population with standard deviation $\sigma = 72$ and unknown mean μ. The sample mean is $\overline{X} = 375.2$. Build a symmetric 68.3% confidence interval for μ.

  Answer: according to the formula, the symmetric 68.3% confidence interval ($k = 1$) for μ would be

  $$\text{C.I.}(\mu; 0.683) \equiv \overline{X} \pm k\frac{\sigma}{\sqrt{n}} \equiv 375.2 \pm 1 \cdot \frac{72}{\sqrt{64}},$$

  which is to say

  $$\text{C.I.}(\mu; 0.683) \equiv (375.2-9, 375.2+9) = (366.2, 384.2).$$

  VERY IMPORTANT: this does not say that we are 68.3% sure that the true μ is between 366.2 and 384.2. Rather, what it says is that when a sample of size 64 is taken from a normal population $\mathcal{N}(\mu, 72^2)$ and a symmetric 68.3% confidence interval for μ is built, μ will fall between the endpoints of the interval about 68.3% of the time.[28]

- Build a symmetric 95.5% confidence interval for μ.

  Answer: the same formula applies, with $k = 2$:

  $$\text{C.I.}(\mu; 0.955) \equiv \overline{X} \pm k\frac{\sigma}{\sqrt{n}} \equiv 375.2 \pm 2 \cdot \frac{72}{\sqrt{64}},$$

  which is to say

  $$\begin{aligned}\text{C.I.}(\mu; 0.995) &\equiv (375.2 - 18, 375.2 + 18) \\ &= (357.2, 393.2).\end{aligned}$$

- Build a symmetric 99.7% confidence interval for μ.

  Answer: again, the same formula applies, with $k = 3$:

  $$\text{C.I.}(\mu; 0.997) \equiv \overline{X} \pm k\frac{\sigma}{\sqrt{n}} \equiv 375.2 \pm 3 \cdot \frac{72}{\sqrt{64}},$$

  which is to say

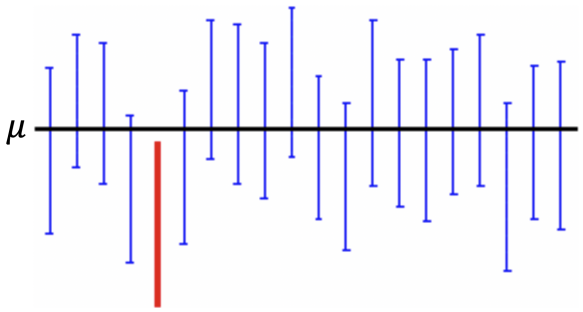
  $$\begin{aligned}\text{C.I.}(\mu; 0.995) &\equiv (375.2 - 27, 375.2 + 27) \\ &= (348.2, 402.2).\end{aligned}$$

Note that the C.I. increases in size with the **confidence level**.

[28] This less than intuitive interpretation of the confidence interval is one of the disadvantages of using the frequentist approach; the analogous concept in Bayesian statistics is called the **credible interval**, which agrees with our naïve expectation of a confidence interval as saying something about how certain we are that the true parameter is in the interval [3, 20].

The interpretation stays the same, no matter the required confidence level or the parameter of interest.

A 95.5% C.I. for the mean, for instance, indicates that we would expect 19 out of 20 samples from the same population to produce confidence intervals that contain the true population mean, **on average**.



Confidence Interval for μ when σ is Known (Reprise) Another approach to C.I. building is to specify the **proportion of the area under $\phi(z)$ of interest**, and then to determine the **critical values** (which is to say, the endpoints of the interval).

Let $\{X_1, \ldots, X_n\}$ be drawn from $N(\mu, \sigma^2)$. Recall that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

For a **symmetric** 95% **C.I. for** μ, we need to find $z^* > 0$ such that $P(-z^* < Z < z^*) \approx 0.95$. But the left-hand side of this "equality" can be re-written as

$$\begin{aligned}P(-z^* < Z < z^*) &= \Phi(z^*) - \Phi(-z^*) \\ &= \Phi(z^*) - (1 - \Phi(z^*)) \\ &= 2\Phi(z^*) - 1;\end{aligned}$$

we are thus looking for a critical value z^* such that

$$0.95 = 2\Phi(z^*) - 1 \implies \Phi(z^*) = \frac{0.95 + 1}{2} = 0.975.$$

From any normal table (or *via* `qnorm(0.975)` in `R`), we see that $\Phi(1.96) \approx 0.9750$, so that

$$P(-1.96 < Z < 1.96) = P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \approx 0.95.$$

In other words, the inequality

$$-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96$$

holds with probability 0.95, or, equivalently,

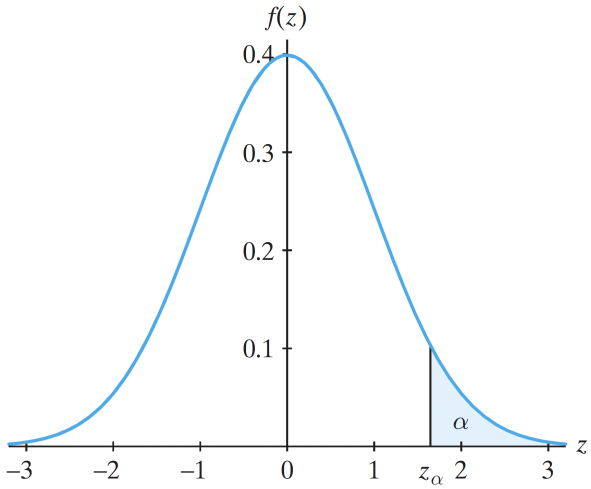
$$\text{C.I.}(\mu; 0.95) \equiv \overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

is the **(symmetric)** 95% **C.I. for** μ **when** σ **is known**.

A similar argument shows that

$$\text{C.I.}(\mu; 0.99) \equiv \overline{X} \pm 2.575 \frac{\sigma}{\sqrt{n}}$$

is the **(symmetric)** 99% **C.I. for μ when σ is known**.



$$P(Z > z_\alpha) = \alpha$$

$$P(Z > z) = 1 - \Phi(z) = \Phi(-z)$$

The **confidence level** $1-\alpha$ is usually expressed in terms of a **small** α, so that $\alpha = 0.05$ corresponds to a confidence level of $1-\alpha = 0.95$.

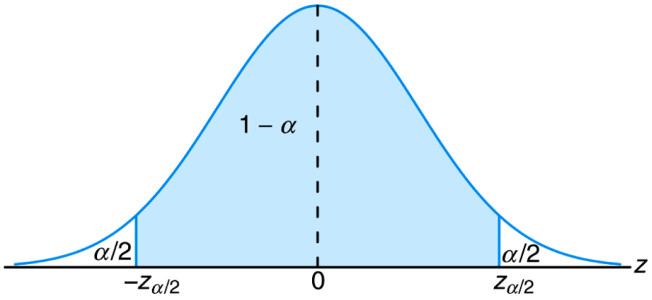
For $\alpha \in (0,1)$, the value z_α for which $P(Z > z_\alpha) = \alpha$ is called the $100(1-\alpha)\%$ **quantiles** of the standard normal distribution.

For general 2−**sided confidence intervals** (the ones we have been building so far), the appropriate numbers are found by solving $P(|Z| > z^*) = \alpha$ for z^*. By the properties of $\mathcal{N}(0,1)$,

$$\begin{aligned}\alpha = P(|Z| > z^*) &= 1 - P(-z^* < Z < z^*) \\ &= 1 - (2\Phi(z^*) - 1) \\ &= 2(1 - \Phi(z^*)),\end{aligned}$$

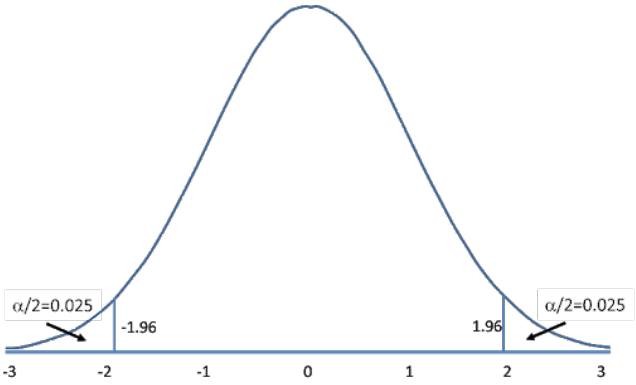
so that

$$\Phi(z^*) = 1 - \alpha/2 \implies z^* = z_{\alpha/2}.$$



For instance,

$$P(|Z| > z_{0.025}) = 0.05 \implies z_{0.025} = 1.96$$
$$P(|Z| > z_{0.005}) = 0.01 \implies z_{0.005} = 2.575.$$



The symmetric $100(1-\alpha)\%$ C.I. for μ can thus generally be written as

$$\text{C.I.}(\mu; 1-\alpha)\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

For a given confidence level α, **shorter confidence intervals are better** in relation to estimating the mean:

- estimates improve when the sample size n increases;
- estimates improve when σ decreases.

For a given sample, if $\alpha_1 > \alpha_2$ then

$$100(1-\alpha_1)\%\ \text{C.I.} \subseteq 100(1-\alpha_2)\%\ \text{C.I.}$$

For instance, the 95% C.I. built from a sample is always contained in the corresponding 99% C.I.

If the sample comes from a normal population, then the C.I. is **exact**. Otherwise, if n is large, we may use the CLT and get an **approximate** C.I.

Examples

- A sample of 9 observations from a normal population with known standard deviation $\sigma = 5$ yields a sample mean $\overline{X} = 19.93$. Provide a 95% and a 99% C.I. for the unknown population mean μ.

  Answer: the point estimate of μ is the sample mean $\overline{X} = 19.93$. The $100(1-\alpha)\%$ C.I.s are

  $$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

  Thus,

  $$\text{C.I.}(\mu; 0.95) \equiv 19.93 \pm 1.96\frac{5}{\sqrt{9}} = (16.66, 23.20)$$
  $$\text{C.I.}(\mu; 0.99) \equiv 19.93 \pm 2.575\frac{5}{\sqrt{9}} = (15.64, 24.22).$$

- A sample of 25 observations from a normal population with known standard deviation $\sigma = 5$ yields a sample mean $\overline{X} = 19.93$. Provide a 95% and a 99% C.I. for the unknown population mean μ.

  Answer: the point estimate of μ is the sample mean $\overline{X} = 19.93$. The $100(1-\alpha)$% C.I.s are

$$\text{C.I.}(\mu; 0.95) \equiv 19.93 \pm 1.96\frac{5}{\sqrt{25}} = (17.97, 21.89)$$
$$\text{C.I.}(\mu; 0.99) \equiv 19.93 \pm 2.575\frac{5}{\sqrt{25}} = (17.35, 22.51).$$

- A sample of 25 observations from a normal population with known standard deviation $\sigma = 10$ yields a sample mean $\overline{X} = 19.93$. Provide a 95% and a 99% C.I. for the unknown population mean μ.

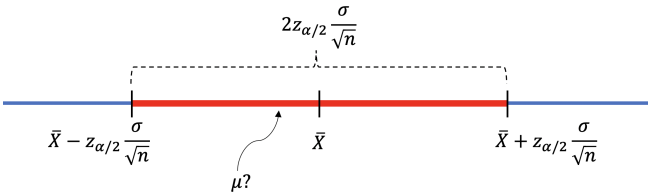
  Answer: the point estimate of μ is the sample mean $\overline{X} = 19.93$. The $100(1-\alpha)$% C.I.s are

$$\text{C.I.}(\mu; 0.95) \equiv 19.93 \pm 1.96\frac{10}{\sqrt{25}} = (16.01, 23.85)$$
$$\text{C.I.}(\mu; 0.99) \equiv 19.93 \pm 2.575\frac{10}{\sqrt{25}} = (14.78, 25.08).$$

Note how the confidence intervals are affected by α, n, and σ.

7.3 Choice of Sample Size

The **error** E we commit by estimating μ *via* the sample mean $\overline{X}$ is smaller than $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, with probability $100(1-\alpha)$% (in the frequentist interpretation).



At this stage, if we want to **control the error** E, the only thing we can really do is control the sample size:[29]

$$E > z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \implies n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.$$

Examples

- A sample $\{X_1, \ldots, X_n\}$ is selected from a normal population with standard deviation $\sigma = 100$. What sample size should be used to insure that the error on the population estimate is at most $E = 10$, at a confidence level $\alpha = 0.05$?

  Answer: as long as

$$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.025}\cdot 100}{10}\right)^2 = (19.6)^2 = 384.16,$$

  then the error committed by using $\overline{X}$ to estimate μ will be at most 10, with 95% probability.

[29] Sampling strategies can also help, but this is a topic for another report.

- Repeat the first example, but with $\sigma = 10$.

  Answer: we need

$$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.025}\cdot 10}{10}\right)^2 = (1.96)^2 = 3.8416.$$

- Repeat the first example, but with $E = 1$.

  Answer: we need

$$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.025}\cdot 100}{1}\right)^2 = (196)^2 = 38416.$$

- Repeat the first example, but with $\alpha = 0.01$.

  Answer: we need

$$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.005}\cdot 100}{10}\right)^2 = (25.75)^2 = 663.0625.$$

The relationship between α, σ, E, and n is not always intuitive, but it follows a simple rule.

7.4 Confidence Interval for μ when σ is Unknown

So far, we have been in the fortunate situation of sampling from a population with **known** variance σ^2. What do we do when the population variance is **unknown** (a situation which occurs much more frequently in real world applications).

The solution is to estimate σ using the **sample variance**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

and the **sample standard deviation** $S = \sqrt{S^2}$; we use $\overline{X}$ instead of μ since we do not know the value of the latter (that is indeed the parameter whose value we are trying to estimate in the first place).[30]

If σ is unknown, it can be shown that $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ follows approximately the **Student** t**−distribution with** $n-1$ **degrees of freedom**, $t(n-1)$.

Consequently, at a confidence level α, we have

$$P\left(-t_{\alpha/2}(n-1) < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1)\right) \approx 1-\alpha,$$

where $t_{\alpha/2}(n-1)$ is the $100(1-\alpha/2)^{\text{th}}$ quantile of $t(n-1)$.[31]

$$100(1-\alpha)\%\text{ C.I. for } \mu \approx \overline{X} \pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}.$$

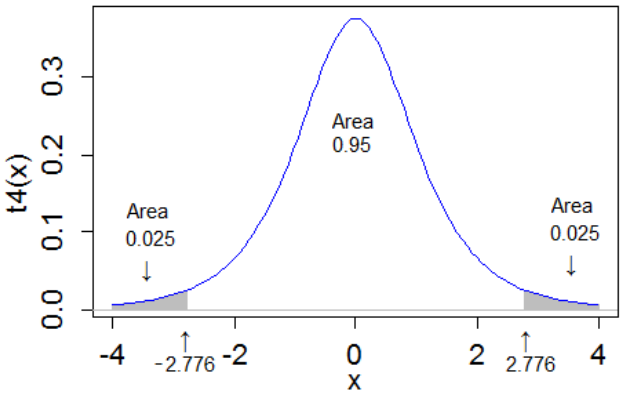
Equality is reached if the underlying population is normal.

[30] Remember, when σ is known (and n is large enough), we already know from the CLT that $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is approximately $\mathcal{N}(0,1)$.

[31] Read from a table or computed using the `R` function `qt()`.

For instance, if $\alpha = 0.05$ and $\{X_1, X_2, X_3, X_4, X_5\}$ are samples from a normal distribution with unknown mean μ and unknown variance σ^2, then $t_{0.025}(5-1) = 2.776$ and

$$P\left(-2.776 < \frac{\overline{X}-\mu}{S/\sqrt{5}} < 2.776\right) = 0.95.$$



Examples

- For a given year, 9 measurements of ozone concentration are obtained:

  $$3.5, 5.1, 6.6, 6.0, 4.2, 4.4, 5.3, 5.6, 4.4.$$

  Assume that the measured ozone concentrations follow a normal distribution with variance $\sigma^2 = 1.21$, build a 95% C.I. for the population mean μ. Note that $\overline{X} = 5.01$ and that $S = 0.97$.

  Answer: since the variance is known, we use the standard normal quantile $z_{\alpha/2} = z_{0.025} = 1.96$:

  $$\overline{X} \pm z_{0.025}\frac{\sigma}{\sqrt{n}} = 5.01 \pm 1.96\frac{\sqrt{1.21}}{\sqrt{9}} = (4.29, 5.73).$$

- Do the same thing, this time assuming that the true variance of the underlying population is unknown.

  Answer: since variance is unknown, we use the Student quantile $t_{\alpha/2}(n-1) = t_{0.025}(8) = 2.306$:

  $$\overline{X} \pm t_{0.025}(n-1)\frac{S}{\sqrt{n}} = 5.01 \pm 2.306\frac{0.97}{\sqrt{9}} = (4.26, 5.76).$$

When the underlying variance is known, the C.I. is **tighter** (smaller), which is only natural as we are more confident about our results when we have more information.

Note: we have seen that when the underlying distribution is normal, or when it is not normal but the sample size is "large" enough, we can build a C.I. for the population mean, whether the population variance is known or not.

If, however, the underlying population is not normal and the sample size is "small", the approach used in this section cannot guarantee the C.I.'s accuracy.

7.5 Confidence Interval for a Proportion

If X is the number of successes in n independent trials, then $X \sim \mathcal{B}(n, p)$, $\mathrm{E}[X] = np$ and $\mathrm{Var}[X] = np(1-p)$, and the point estimator for p is $\hat{P} = \frac{X}{n}$.

Since X is a sum of iid random variables, its **standardization**

$$Z = \frac{X-\mu}{\sigma} = \frac{n\hat{P}-np}{\sqrt{np(1-p)}} = \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately $\mathcal{N}(0,1)$, when n is large enough.

Thus, for sufficiently large n,

$$P\left(-z_{\alpha/2} < \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) \approx 1-\alpha.$$

Using the construction presented earlier in this section, we conclude that

$$\hat{P} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

is an **approximate** $100(1-\alpha)\%$ C.I. for p. However, this result is not useful in practice because p is unknown, so we use the following approximation instead:

$$\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

Examples

- Two candidates (A and B) are running for office. A poll is conducted: 1000 voters are selected randomly and asked for their preference: 52% support A, while 48% support their rival, B. Provide a 95% C.I. for the support of each candidate.

  Answer: we use $\alpha = 0.05$ and $\hat{P} = 0.52$. The approximate 95% C.I. for A is thus

  $$0.52 \pm 1.96\sqrt{\frac{0.52 \cdot 0.48}{1000}} \approx 0.52 \pm 0.031,$$

  while the one for B is 0.48 ± 0.031.

- On the strength of this polling result, a newspaper prints the following headline: "Candidate A Leads Candidate B!" Is the headline warranted?

  Answer: although there is a 4—point gap in the poll numbers, the true support for candidate A is in the 48.9%—55.1% range, and, the true support for candidate B is in the 44.9%—51.1% range, with probability 95% (that is to say, 19 times out of 20).

  Since there is overlap in the confidence intervals, the race is more likely to be a dead heat.

8. Hypothesis Testing

Hypothesis testing is another of the myriads of statistical analysis concerns.

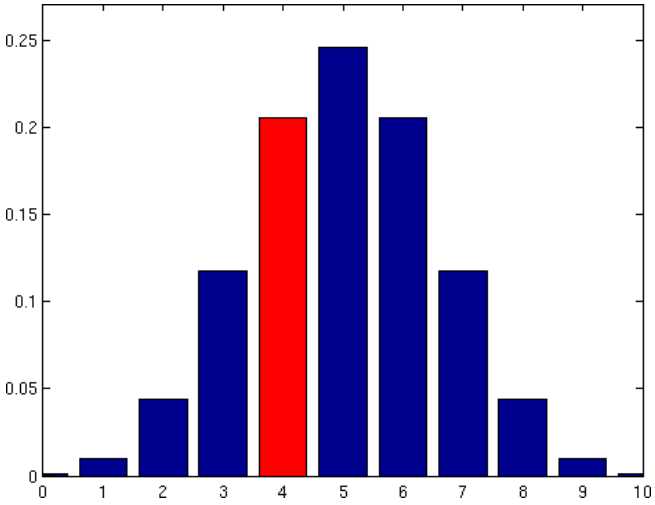
Claims and Suspicions Consider the following scenario: person A claims they have a fair coin, but for some reason, person B is suspicious of the claim, believing the coin to be biased in favour of tails.

Person B flips the coin 10 times, expecting a low number of heads, which they intend to use as **evidence** against the claim. Let $X = \#$ of Heads.

Suppose $X = 4$. This is less than expected for a binomial random variable $X \sim \mathscr{B}(10, 0.5)$ since $\mathrm{E}[X] = 5$; the results are more in line with a coin for which $P(\text{Head}) = 0.4$.

Does this data really constitute evidence against the claim $P(\text{Head}) = 0.5$?

If the coin is fair, then $X \sim \mathscr{B}(10, 0.5)$ and $X = 4$ is still close to $\mathrm{E}[X]$; in fact, $P(X = 4) = 0.205$ (as opposed to $P(X = 5) = 0.246$) so the event $X = 4$ is still quite likely. It would seem that there is no *real* evidence against the claim that the coin is fair.



The way the sentence "*It would seem that there is no evidence against the claim that the coin is fair*" is worded is very important.

We did not reject the claim that $P(\text{Head}) = 0.5$ (i.e. that the coin is symmetric), but this **doesn't mean that, in fact,** $P(\textbf{Head}) = 0.5$.

Not rejecting (which is not quite the same as "accepting") a claim is a **very weak statement**.

To see why, let's consider person C, who claims that the coin from the example above has $P(\text{Head}) = 0.3$. Under $X \sim \mathscr{B}(10, 0.3)$, the event $X = 4$ is still quite likely, with $P(X = 4) = 0.22$; we **do not have enough evidence to reject** either $P(\text{Head}) = 0.5$ or $P(\text{Head}) = 0.3$.

However, **rejecting** a claim is a **very strong statement**!

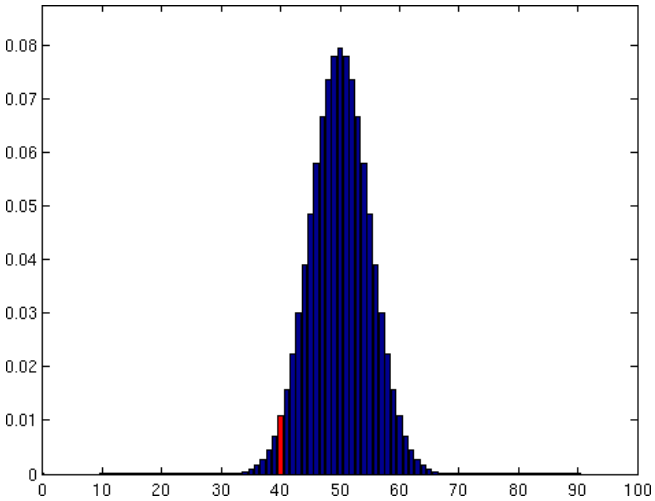
Let's say that person B convinces person A to flip the coin another 90 times. In the second round of flips, 36 Heads occur, giving a total of 40 Heads out of 100 coin flips.

What can we say now? Does this constitute any evidence against the claim? If so, how much?

Let $Y \sim \mathscr{B}(100, 0.5)$ (i.e. the coin is fair); $Y = 40$ is smaller than what we would expect as $\mathrm{E}[Y] = 50$ if the claim is true, so $Y = 40$ is again more in agreement with $P(\text{Head}) = 0.4$.

But the event $Y = 40$ **does not** lie in the probability mass centre of the distribution; it falls in the **distribution tail** (an area of lower probability).

For $Y \sim \mathscr{B}(100, 0.5)$, $P(Y = 40) = 0.011$ (compare this with the previous value 0.205). Thus, if the coin is fair, the event $Y = 40$ is quite **unlikely**.



Values down in the lower tail (or up in the upper tail) provide **some evidence** against the claim. The question is, how much evidence? **How do we quantify it?**

Since values that are "further down the left tail" provide evidence against the claim of a fair coin (in favour of a coin biased against Heads), we will use the actual tail area that goes with the observation: **the smaller the tail area, the greater the evidence against the claim** (and *vice-versa*).

For 4 Heads out of 10 tosses, the evidence is the $p-$**value** $P(X \leq 4)$ if the claim is true, i.e.

$$P(X \leq 4 \mid X \sim \mathscr{B}(10, 0.5)) = 0.377.$$

Thus, if $P(\text{Head}) = 0.5$, the event $X \leq 4$ is still very likely: we would see evidence that extreme (or more) $\approx 38\%$ of the time (simply by chance).

For 40 Heads out of 100 tosses, the evidence is the $p-$**value** $P(Y \leq 40)$ if the claim is true, i.e.

$$P(Y \leq 40 \mid Y \sim \mathscr{B}(100, 0.5)) = 0.028.$$

Thus, if $P(\text{Head}) = 0.5$, the event $Y \leq 40$ is very unlikely: we would only see evidence that extreme (or more) $\approx 3\%$ of the time.

A claim's $p-$value is the **area of the tail** of the distribution's p.d.f. under the assumption that the claim is true:

$$\text{smaller } p-\text{value} \iff \text{more evidence against claim.}$$

A specific language and notation has evolved to describe this approach to "testing hypotheses":

- the "claim" is called the **null hypothesis** and is denoted by H_0.
- the "suspicion" is called the **alternative hypothesis** and is denoted by H_1;
- the (random) quantity we use to measure evidence is called a **test statistic** – we need to know its distribution when H_0 is true, and
- the $p-$**value** quantifies "the evidence against H_0".

Consider the coin tossing situation described previously. The null hypothesis is

$$H_0 : P(\text{Head}) = 0.5\,.$$

The alternative hypothesis is

$$H_1 : P(\text{Head}) < 0.5\,.$$

The coin is tossed n times; the test statistic is the number of heads X in n tosses.

- If $n = 10$ and $X = 4$, the $p-$value is

$$P(X \leq 4 \mid X \sim \mathscr{B}(10, 0.5)) = 0.377,$$

  on the basis of which we would not reject the null hypothesis that the coin was fair.

- If $n = 100$ and $X = 40$, the $p-$value is

$$P(X \leq 40 \mid X \sim \mathscr{B}(100, 0.5)) = 0.028,$$

  on the basis of which we would reject the null hypothesis that the coin was fair, in favour of the alternative that it was not.

How Small Does the $p-$Value Need to Be? We concluded that 37.7% was "not that small", whereas 2.8% was "small enough;. How small does a $p-$value need to be before we consider that we have "compelling evidence" against H_0?

There is no easy answer to this question. It depends on many factors, including what penalties we might pay for being wrong.

Typically, we look at the probability of making a **type I error**, $\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$:

- if $p-$value $\leq \alpha$, then we **reject** H_0 in favour of H_1;
- if $p-$value $> \alpha$, then **there is not enough evidence to reject** H_0 (**which is not the same as accepting** H_0).

By convention, we often use $\alpha = 0.01$ or $\alpha = 0.05$.

The use of p-values has come under fire recently, as many view them as the root cause of the current **replication crisis**.[32] In this twitter thread , K. Carr describes why there isn't something wrong with $p-$values *per se*:

> Don't know what a P-VALUE is?
> Don't know why P-VALUES work?
> Don't know why sometimes P-VALUES don't work?
> THIS IS THE THREAD FOR YOU.
> DEFINITION OF A P-VALUE. Assume your theory is false. The P-VALUE is the probability of getting an outcome as extreme or even more extreme than what you got in your experiment.
> THE LOGIC OF THE P-VALUE. Assume my theory is false. The probability of getting extreme results should be very small but I got an extreme result in my experiment. Therefore, I conclude that this is strong evidence that my theory is true. That's the logic of the p-value.
> THE P-VALUE IS REASONABLE IN THEORY BUT TRICKY IN PRACTICE. In my opinion, the p-value is just a mathematical version of the way humans think. If we see something that seems unlikely given our beliefs, we often doubt those beliefs. In practice, the p-value can be tricky to use.
> THE P-VALUE REQUIRES A GOOD DEFINITION OF WHEN YOUR THEORY IS FALSE. There are usually an infinite number of ways to define a world where your theory is false. P-values often fail when people use overly simplistic mathematical models of the processes that created their data. If the mismatch between their mathematical models of the world and the actual world is too large then the probabilities we compute can become completely disconnected from reality.
> THE P-VALUE MAY REQUIRE AN ACCURATE MODEL OF YOU (THE OBSERVER). The probability of getting the result you got depends on many things. If you sometimes do things like throw out data or repeat measurements then you're part of the system. Your behavior affects the probability of getting your experimental results. Therefore, to be completely realistic, you need to have an ACCURATE model of your own behavior when you gather and analyze data. This is hard and a big part of why the p-value often fails as a tool.
> BY DEFINITION, P-VALUES MUST SOMETIMES BE WRONG. When using p-values, we're working off of probabilities. By logic of the p-value itself, even with perfect use, some of your decisions will be wrong. You have to embrace this if you're going to use the p-values. Badly defining what it means for your model to be false. Inaccurately modeling the chances of getting your data including your own behaviors. Not treating a p-value as a decision rule that can sometimes be wrong. These factors all contribute to misuse of the p-value in practice. Hope this cleared some things up for you.
> Thanks for coming to my p-value TED talk!

[32] The crisis concerns the prevalence of positive findings that are contradicted in subsequent studies [4].

8.1 Hypothesis Testing

A **hypothesis** is a conjecture concerning the value of a population parameter.

Hypothesis testing require two competing hypotheses:

- a **null hypothesis**, denoted by H_0;
- an **alternative hypothesis**, denoted by H_1 or H_A.

The hypothesis is **tested** by evaluating experimental evidence:

- if the evidence against H_0 is **strong enough**, we reject H_0 **in favour of** H_1, and we say that the evidence against H_0 in favour of H_1 is **significant**;
- if the evidence against H_0 is **not** strong enough, then we fail to reject H_0 and we say that the evidence against H_0 is **not significant**.

In cases when we fail to reject H_0, we **do NOT accept** H_0 instead – we simply do not have enough evidence to reject H_0.

The hypotheses should be formulated **prior to the experiment** or the study. The experiment or study is then conducted to evaluate the evidence against the null hypothesis – in order to avoid **data snooping**, it is crucial that we do not formulate H_1 after looking at the data.

Scientific hypotheses can be often expressed in terms of whether an effect is found in the data.

In this case, we use the following null hypothesis:

$$H_0 : \text{there is no effect}$$

against the alternative hypothesis:

$$H_1 : \text{there is an effect.}$$

Errors in Hypothesis Testing Two types of errors can be committed when testing H_0 against H_1

- If we reject H_0 when H_0 was in fact true, we have committed a **type I error**;
- if we fail to reject H_0 when H_0 was in fact is false, we have committed a **type II error**.

	Decision: reject H_0	**Decision:** fail to reject H_0
Reality: H_0 is True	Type I Error	No Error
Reality: H_0 is False	No Error	Type II Error

Examples

- If we conclude that a drug treatment is useful for treating a particular disease, but this is not the case in reality, then we have committed an error of type I.
- If we cannot conclude that a drug treatment is useful for treating a particular disease, but in reality the treatment is effective, then we have committed an error of type II.

What type of error is worst? It depends on many factors.

Power of a Test The probability of committing a type I error is usually denoted by

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true});$$

that of committing a type II error by

$$\beta = P(\text{fail to reject } H_0 \mid H_0 \text{ is false}),$$

and that of correctly rejecting H_0 by

$$\text{power} = P(\text{reject } H_0 \mid H_0 \text{ is false}) = 1 - \beta.$$

Conventional values of α and β are usually 0.05 and 0.2, respectively, although that is not a hard rule.

Types of Null and Alternative Hypotheses Let μ be the population parameter of interest. The hypotheses are expressed in terms of the values of this parameter.

The null hypothesis is a **simple hypothesis** of the form:

$$H_0 : \mu = \mu_0,$$

where μ_0 is some candidate value ("simple" means that it is assumed to be a single value.)

The alternative hypothesis H_1 is a **composite hypothesis**, i.e. it contains more than one candidate value.

Depending on the context, hypothesis testing takes on one of the following three forms:

$$H_0 : \mu = \mu_0, \quad \text{where } \mu_0 \text{ is a number,}$$

against a:

- **two-sided** alternative: $H_1 : \mu \neq \mu_0$;
- **left-sided** alternative: $H_1 : \mu < \mu_0$, or
- **right-sided** alternative: $H_1 : \mu > \mu_0$.

The formulation of the alternative hypothesis depends on the research hypothesis and is determined **prior** to experiment or study.

Example: investigators often want to verify if new experimental conditions lead to a change in population parameters.

For instance, an investigator claims that the use of a new type of soil will produce taller plants on average compared to the use of traditional soil. The mean plant height under the use of traditional soil is 20 cm.

1. Formulate the hypotheses to be tested.
2. If another investigator suspects the opposite, that is, that the mean plant height when using the new soil will be smaller than the mean plant height with old soil. What hypotheses should be formulated?
3. A 3rd investigator believes that there will be an effect, but is not sure if the effect with be to produce shorter or taller plants. What hypotheses should be formulated then?

Answer: let μ represent the mean plant height with the new type of soil. In all three cases, the null hypothesis is $H_0 : \mu = 20$.

The alternative hypothesis depends on the situation:

1. $H_1 : \mu > 20$.
2. $H_1 : \mu < 20$.
3. $H_1 : \mu \neq 20$.

For each H_1, the corresponding $p-$values would be computed differently when testing H_0 against H_1.

8.2 Test Statistics and Critical Regions

We test a statistical hypothesis we use a **test statistic**. A test statistic is a function of the random sample and the population parameter of interest.

In general, we reject H_0 if the value of the test statistic is in the **critical region** or **rejection area** for the test; the critical region is an interval of real numbers.

The critical region is obtained using the definition of errors in hypothesis testing – we select the critical region so that

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

is equal to some pre-determined value, such as 0.05 or 0.01.

Examples: a new curing process developed for a certain type of cement results in a mean compressive strength of 5000 kg/cm^2, with a standard deviation of 120 kg/cm^2.

We test the hypothesis $H_0 : \mu = 5000$ against the alternative $H_1 : \mu < 5000$ with a random sample of 49 pieces of cement. Assume that the critical region in this specific instance is $\overline{X} < 4970$, that is, we would reject H_0 if $\overline{X} < 4970$.

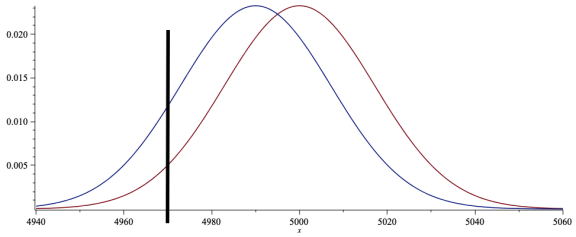
- Find the probability of committing a type I error when H_0 is true.

  Answer: by definition, we have

  $$\begin{aligned}\alpha &= P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true}) \\ &= P(\overline{X} < 4970 \mid \mu = 5000).\end{aligned}$$

  Thus, according to the CLT, we have

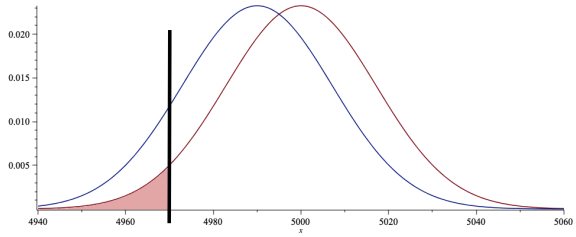
  $$\begin{aligned}\alpha &\approx P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{4970-5000}{120/7}\right) \\ &\approx P(Z < -1.75) \approx 0.0401\,.\end{aligned}$$



The sampling distribution of $\overline{X}$ under H_0 is shown in **red** in the graphs (mean = 5000, sd = 120/7); the sampling distribution of $\overline{X}$ under H_1 in **blue** (mean = 4990, sd = 120/7).

The critical region falls to the left of the vertical **black** line $\overline{X} < 4970$, and the probability of committing a type I error is the area shaded in red:

$$\begin{aligned}\alpha &= P(\text{reject } H_0 \mid H_0 \text{ is true}) \\ &= P(\overline{X} < 4970 \mid \mu = 5000).\end{aligned}$$



We would thus reject H_0 if the observed value of $\overline{X}$ falls to the left of $\overline{X} = 4970$ (in the critical region).

- Evaluate the probability of committing a type II error if μ is actually 4990, say (and not 5000, as in H_0).

  Answer: by definition, we have

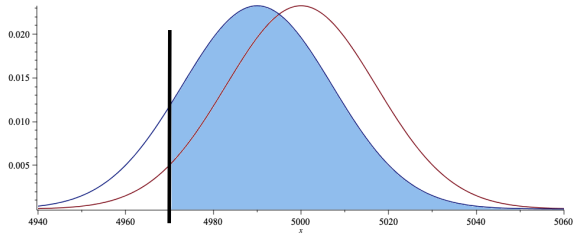
  $$\begin{aligned}\beta &= P(\text{type II error}) = P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) \\ &= P(\overline{X} > 4970 \mid \mu = 4990).\end{aligned}$$

  Thus, according to the CLT, we have

  $$\begin{aligned}\beta &= P(\overline{X} > 4970) = P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{4970-4990}{120/7}\right) \\ &\approx P(Z > -1.17) = 1 - P(Z < -1.17) \approx 0.879\,.\end{aligned}$$

  The critical region falls to the the right of the vertical black line, and the probability of committing a type II error is the area shaded in blue:

  $$\begin{aligned}\beta &= P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) \\ &= P(\overline{X} > 4970 \mid \mu = 4990).\end{aligned}$$

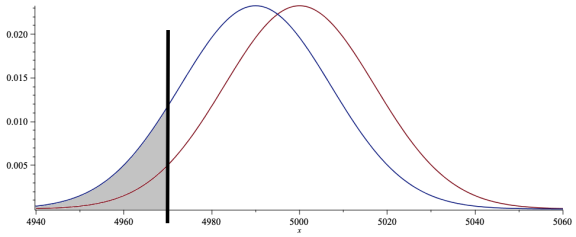


We would thus fail to reject H_0 if the observed vale of $\overline{X}$ falls to the right of $\overline{X} = 4970$ (outside the critical region).

The power of the test is easily computed as

$$\begin{aligned}\text{power} &= P(\text{reject } H_0 \mid H_0 \text{ is false}) \\ &= P(\overline{X} < 4970) = 1-\beta \approx 0.121,\end{aligned}$$

the area shaded in grey.



- Evaluate the probability of committing a type II error if μ is actually 4950, say (and not 5000, as in H_0).

  Answer: by definition, we have

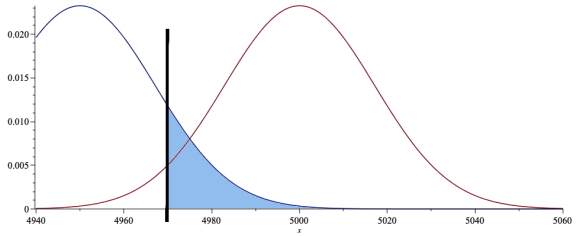
$$\begin{aligned}\beta &= P(\text{type II error}) \\ &= P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) \\ &= P(\overline{X} > 4970 | \mu = 4950).\end{aligned}$$

  Thus, according to the CLT, we have

$$\begin{aligned}\beta &= P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{4970-4950}{120/7}\right) \\ &\approx P(Z > 1.17) \approx 0.121\,.\end{aligned}$$

  The critical region falls to the the right of the vertical black line, and the probability of committing a type II error is the area shaded in blue:

$$\begin{aligned}\beta &= P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) \\ &= P(\overline{X} > 4970 \mid \mu = 4950).\end{aligned}$$



  We would thus fail to reject H_0 if the observed value of $\overline{X}$ falls to the right of $\overline{X} = 4970$ (outside the critical region).

The probability of making a type II error is substantially larger in the first case, which means that the threshold $\overline{X} = 4970$ is not ideal in that situation.

8.3 Test for a Mean

Suppose $X_1, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2, and let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote the sample mean. We have seen that

- if the population is normal, then $\overline{X} \overset{\text{exact}}{\sim} \mathcal{N}(\mu, \sigma^2/n)$;
- if the population is **not** normal, then as long as n is **large enough**, $\overline{X} \overset{\text{approx}}{\sim} \mathcal{N}(\mu, \sigma^2/n)$.

In this section, we start by assuming that the population variance σ^2 is **known**, and that the hypothesis concerns the **unknown** population mean μ.

Explanation: Left-Sided Alternative Consider the unknown population mean μ. Suppose that we would like to test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu < \mu_0,$$

where μ_0 is some candidate value for μ.

To evaluate the evidence against H_0, we compare $\overline{X}$ to μ_0. Under H_0,

$$Z_0 = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1).$$

We say that $z_0 = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$ is the observed value of the Z**—test statistic** Z_0.

If $z_0 < 0$, we have evidence that $\mu < \mu_0$. However, we only reject H_0 in favour of H_1 if the evidence is **significant**, which is to say, if

$$z_0 \leq -z_\alpha, \text{ at a level of significance } \alpha.$$

The corresponding p**—value** for this test is the probability of observing evidence as or more extreme than our current evidence in favour of H_1, assuming that H_0 is true (that is, simply by chance).[33]

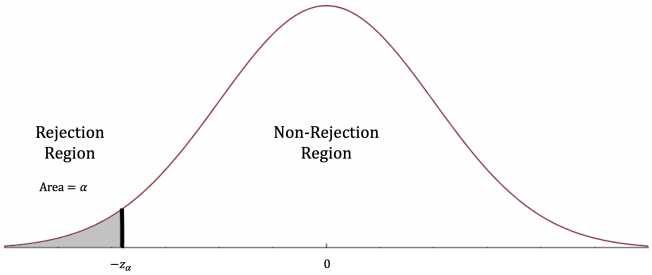
The **decision rule** for the left-sided test is thus

- if the p—value $\leq \alpha$, we **reject** H_0 **in favour of** H_1;
- if the p—value $> \alpha$, we **fail to reject** H_0**.**

Formally, the **left-sided test** pits

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu < \mu_0;$$

at significance α, if $z_0 = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha$, we reject H_0 in favour of H_1.

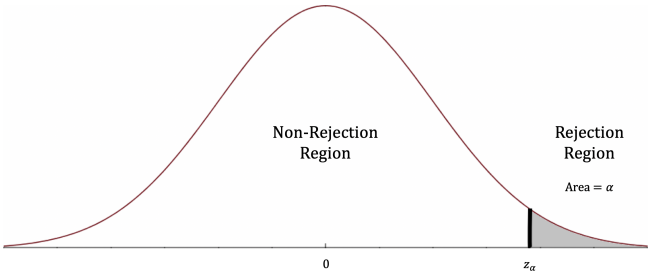


[33] "Even more extreme", in this case, means further to the left, so that p-value $= P(Z \leq z_0) = \Phi(z_0)$, where z_0 is the observed value for the Z-test statistic.

An equivalent **right-sided test** pits

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu > \mu_0;$$

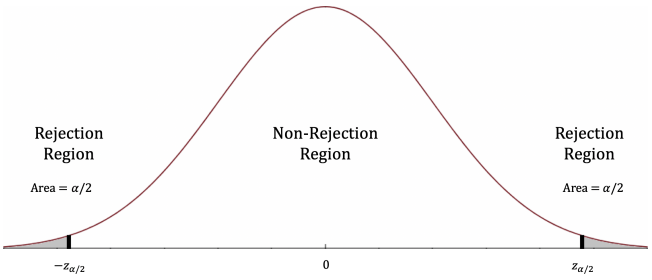
at significance α, if $z_0 = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$, we reject H_0 in favour of H_1.



The **two-sided test** pits

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0;$$

at significance α, if $|z_0| = \left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right| \geq z_{\alpha/2}$, we reject H_0 in favour of H_1.



The **procedure** to test for $H_0 : \mu = \mu_0$ requires 6 steps.

Step 1: set $H_0 : \mu = \mu_0$.

Step 2: select an alternative hypothesis H_1 (what we are trying to show using the data). Depending on the context, we choose one of these alternatives:

- $H_1 : \mu < \mu_0$ (one-sided test);
- $H_1 : \mu > \mu_0$ (one-sided test);
- $H_1 : \mu \neq \mu_0$ (two-sided test).

Step 3: choose $\alpha = P(\text{type I error})$, typically $\alpha \in \{0.01, 0.05\}$.

Step 4: for the observed sample $\{x_1, \ldots, x_n\}$, compute the observed value of the test statistics $z_0 = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$.

Step 5: determine the critical region according to:

Alternative Hypothesis	Critical Region
$H_1 : \mu > \mu_0$	$z_0 > z_\alpha$
$H_1 : \mu < \mu_0$	$z_0 < -z_\alpha$
$H_1 : \mu \neq \mu_0$	$\lvert z_0 \rvert > z_{\alpha/2}$

where z_α is the critical value satisfying $P(Z > z_\alpha) = \alpha$, for $Z \sim \mathcal{N}(0,1)$.

The critical values are displayed below for convenience.

α	z_α	$z_{\alpha/2}$
0.05	1.645	1.960
0.01	2.327	2.576

Step 6: compute the associated $p-$value according to:

Alt. Hypothesis	Critical Region
$H_1 : \mu > \mu_0$	$P(Z > z_0)$
$H_1 : \mu < \mu_0$	$P(Z < z_0)$
$H_1 : \mu \neq \mu_0$	$2 \cdot \min\{P(Z > z_0), P(Z < z_0)\}$

Decision Rule: as above,

- if the $p-$value $\leq \alpha$, **reject** H_0 **in favour of** H_1;
- if the $p-$value $> \alpha$, **fail to reject** H_0.

A few examples will clarify the procedure.

Examples

- Components are manufactured to have strength normally distributed with mean $\mu = 40$ units and standard deviation $\sigma = 1.2$ units. The manufacturing process has been modified, and an increase in mean strength is claimed (the standard deviation remains the same).

  A random sample of $n = 12$ components produced using the modified process had the following strengths:

  ```
  42.5, 39.8, 40.3, 43.1, 39.6, 41.0,
  39.9, 42.1, 40.7, 41.6, 42.1, 40.8.
  ```

  Does the data provide strong evidence that the mean strength now exceeds 40 units? Use $\alpha = 0.05$.

  Answer: we follow the outlined procedure to test for $H_0 : \mu = 40$ against $H_1 : \mu > 40$.

  The observed value of the sample mean is $\overline{x} = 41.125$. Hence,

$$\begin{aligned} p-\text{value} &= P(\overline{X} \geq \overline{x}) = P(\overline{X} \geq 41.125) \\ &= P\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \geq \frac{41.125-\mu_0}{\sigma/\sqrt{n}}\right) \\ &= P(Z \geq 3.25) \approx 0.006. \end{aligned}$$

  As the $p-$value is smaller than α, we reject H_0 in favour of H_1.

  Another way to see this is that if the model '$\mu = 40$' is true, then it is very unlikely that we would observe the event $\{\overline{X} \geq 41.125\}$ entirely by chance, and so the manufacturing process likely has an effect in the claimed direction.

- A set of scales works properly if the measurements differ from the true weight by a normally distributed random error term with standard deviation $\sigma = 0.007$ grams. Researchers suspect that the scale is systematically adding to the weights.

  To test this hypothesis, $n = 10$ measurements are made on a 1.0g "gold-standard" weight, giving a set of measurements which average out to 1.0038g. Does this provide evidence that the scale adds to the measurement weights? Use $\alpha = 0.05$ and 0.01.

  Answer: let μ be the weight that the scale would record in the absence of random error terms. We test for $H_0 : \mu = 1.0$ against $H_1 : \mu > 1.0$.

  The observed test statistic is $z_0 = \frac{1.0038-1.0}{0.007/\sqrt{10}} \approx 1.7167$. Since

  $$z_{0.05} = 1.645 < z_0 = 1.7167 \leq z_{0.01} = 2.327,$$

  we reject H_0 for $\alpha = 0.05$, but we fail to reject H_0 for $\alpha = 0.01$. Case closed. Right?

- In the previous example, assume that we are interested in whether the scale works properly, which means that the investigators think there might be some systematic misreading, but they are not sure in which direction the misreading would occur. Does the sample data provide evidence that the scale is systematically biased? Use $\alpha = 0.05$ and 0.01.

  Answer: let μ be as in the previous example. We test for $H_0 : \mu = 1.0$ against $H_1 : \mu \neq 1.0$.

  The test statistic is still $z_0 = 1.7167$; since $|z_0| \leq z_{\alpha/2}$ for both $\alpha = 0.05$ and $\alpha = 0.01$, we fail to reject H_0 at either $\alpha = 0.05$ or $\alpha = 0.01$.

  Thus, our "reading" of the test statistic depends on what type of alternative hypothesis we have selected (and so, on the overall context).

- The marks for an "average" class are normally distributed with mean 60 and variance 100. Nine students are selected from the class; their average mark is 55. Is this subgroup "below average"?

  Answer: let μ be the true mean of the subgroup. We are testing for $H_0 : \mu = 60$ against $H_1 : \mu < 60$.

  The observed sample test statistic is

  $$z_0 = \frac{55-60}{10/\sqrt{9}} = -1.5.$$

  The corresponding p—value is

  $$P(\overline{X} \leq 55) = P(Z \leq -1.5) = 0.07.$$

  Thus there is not enough evidence to reject the claim that the subgroup is 'average', regardless of whether we use $\alpha = 0.05$ or $\alpha = 0.01$.

- We consider the same set-up as in the previous example, but this time the sample size is $n = 100$, not 9. Is there some evidence to suggest that this subgroup of students is 'below average'?

  Answer: let μ be as before. We are still testing for $H_0 : \mu = 60$ against $H_1 : \mu < 60$, but this time the observed sample test statistic is

  $$z_0 = \frac{55-60}{10/\sqrt{100}} = -5.$$

  The corresponding p—value is

  $$P(\overline{X} \leq 55) = P(Z \leq -5) \approx 0.00.$$

  Thus we reject the claim that the subgroup is 'average', regardless of whether we use $\alpha = 0.05$ or $\alpha = 0.01$.

The lesson from the last example is that the **sample size plays a role**; in general, an estimate obtained from a larger (representative) sample is more likely to be generalizable to the population as a whole.

Tests and Confidence Intervals It is becoming more and more common for analysts to bypass the computation of the p—value altogether, in favour of a confidence interval based approach.[34]

For a given α, we reject $H_0 : \mu = \mu_0$ in favour of $H_1 : \mu \neq \mu_0$ if, and only if, μ_0 is **not** in the $100(1-\alpha)\%$ C.I. for μ.

Example: A manufacturer claims that a particular type of engine uses 20 gallons of fuel to operate for one hour. It is known from previous studies that this amount is normally distributed with variance $\sigma^2 = 25$ and mean μ.

A sample of size $n = 9$ has been taken and the following value has been observed for the mean amount of fuel per hour: $\overline{X} = 23$. Should we accept the manufacturer's claim? Use $\alpha = 0.05$.

Answer: we test for $H_0 : \mu = 20$ against $H_1 : \mu \neq 20$. The observed sample test statistic is

$$z_0 = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{23-20}{5/\sqrt{9}} = 1.8.$$

For a 2—sided test with $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$. Since $|z_0| \leq z_{0.025}$, z_0 is not in the critical region, and we do not reject H_0.

The advantage of the **confidence interval** approach is that it allows analysts to test for various claims **simultaneously**. Since we know the variance of the underlying population, an approximate $100(1-\alpha)\%$ C.I. for μ is given by

$$\overline{X} \pm z_{\alpha/2}\sigma/\sqrt{n} = 23 \pm 1.96 \cdot 5/\sqrt{9} = (19.73; 26.26).$$

Based on the data, we would thus not reject the claim that $\mu = 20$, $\mu = 19.74$, $\mu = 26.20$, etc.

[34] In order to avoid the controversy surrounding the crisis of replication?

Test for a Mean with Unknown Variance If the data is normal and σ is unknown, we can estimate it *via* the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

As we have seen for confidence intervals, the test statistic

$$T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

follows a **Student's t−distribution with $n-1$ df**.

We can follow the same steps as for the test with known variance, with the modified critical regions and p−values:

Alternative Hypothesis	Critical Region
$H_1 : \mu > \mu_0$	$t_0 > t_\alpha(n-1)$
$H_1 : \mu < \mu_0$	$t_0 < -t_\alpha(n-1)$
$H_1 : \mu \neq \mu_0$	$\lvert t_0\rvert > t_{\alpha/2}(n-1)$

where

$$t_0 = \frac{\overline{x}-\mu_0}{S/\sqrt{n}}$$

and $t_\alpha(n-1)$ is the t−value satisfying

$$P(T > t_\alpha(n-1)) = \alpha$$

for $T \sim t(n-1)$, and

Alt. Hypothesis	p−Value
$H_1 : \mu > \mu_0$	$P(T > t_0)$
$H_1 : \mu < \mu_0$	$P(T < t_0)$
$H_1 : \mu \neq \mu_0$	$2 \cdot \min\{P(T > t_0), P(T < t_0)\}$

Example: consider the following observations, taken from a normal population with unknown mean μ and variance:

```
18.0, 17.4, 15.5, 16.8, 19.0, 17.8,
17.4, 15.8, 17.9, 16.3, 16.9, 18.6,
17.7, 16.4, 18.2, 18.7.
```

Conduct a right-side hypothesis test for $H_0 : \mu = 16.6$ against $H_1 : \mu > 16.6$, using $\alpha = 0.05$.

Answer: the sample size, sample mean, and sample variance are $n = 16$, $\overline{X} = 17.4$ and $S = 1.078$, respectively.

Since the variance σ^2 is unknown, the observed sample test statistics of interest is

$$t_0 = \frac{\overline{x}-\mu_0}{S/\sqrt{n}} = \frac{17.4-16.6}{1.078/4} \approx 2.968,$$

and the corresponding p−value is

$$p\text{−value } = P(\overline{X} \geq 17.4) = P(T > 2.968),$$

where $T \sim t(n-1) = t(\nu) = t(15)$.

From the t−tables (or by using the R function `qt()`), we see that

$$P(T(15) \geq 2.947) \approx 0.005,\ P(T(15) \geq 3.286) \approx 0.0025.$$

The p−value thus lies in the interval $(0.0025, 0.005)$; in particular, the p−value ≤ 0.05, which is strong evidence against $H_0 : \mu = 16.6$.

8.4 Test for a Proportion

The principle for proportions is pretty much the same; as we can see in the next example.

Example: a group of 100 adult American Catholics were asked the following question: "Do you favour allowing women into the priesthood?" 60 of the respondents independently answered 'Yes'; is the evidence strong enough to conclude that more than half of American Catholics favour allowing women to be priests?

Answer: let X be the number of people who answered 'Yes'. We assume that $X \sim \mathcal{B}(100, p)$, where p is the true proportion of American Catholics who favour allowing women to be priests.

We thus test for $H_0 : p = 0.5$ against $H_1 : p > 0.5$. Under H_0, $X \sim \mathcal{B}(100, 0.5)$.

The p−value that corresponds to the observed sample is

$$\begin{aligned} P(X \geq 60) &= 1 - P(X < 60) = 1 - P(X \leq 59) \\ &\approx 1 - P\left(\frac{X+0.5-np}{\sqrt{np(1-p)}} \leq \frac{59+0.5-50}{\sqrt{25}}\right) \\ &\approx 1 - P(Z \leq 1.9) = 0.0287, \end{aligned}$$

where the $+0.5$ comes from the correction to the normal approximation of the binomial distribution (see Section 3.6 for details).

Thus, we would reject H_0 at $\alpha = 0.05$, but not at $\alpha = 0.01$.

8.5 Two-Sample Tests

Up to this point, we have only tested hypotheses about populations by evaluating the evidence provided by a single sample of observations.

Two-sample tests allows analysts to compare two (potentially distinct) populations.

Paired Test Let $X_{1,1}, \ldots, X_{1,n}$ be a random sample from a normal population with unknown mean μ_1 and unknown variance σ^2; let $X_{2,1}, \ldots, X_{2,n}$ be a random sample from a normal population with unknown mean μ_2 and unknown variance σ^2, with both populations **not necessarily independent** of one another (i.e., it's possible that the 2 samples arise from the same population, or represent two different measurements on the same units).

We would like to test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$.

In order to do so, we compute the differences $D_i = X_{1,i} - X_{2,i}$ and consider the $t-$test (as we do not know the variance). The test statistic is

$$T_0 = \frac{\overline{D}}{S_D/\sqrt{n}} \sim t(n-1),$$

where

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \text{ and } S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2.$$

Example: the knowledge of basic statistical concepts for $n = 10$ engineers was measured on a scale from $0-100$ *before* and *after* a short course in statistical quality control. The result are as follows:

Engineer	1	2	3	4	5
Before $X_{1,i}$	43	82	77	39	51
After $X_{2,i}$	51	84	74	48	53
Engineer	6	7	8	9	10
Before $X_{1,i}$	66	55	61	79	43
After $X_{2,i}$	61	59	75	82	48

Let μ_1 and μ_2 be the mean score before and after the course, respectively.

Assuming the underlying scores are normally distributed, test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$.

Answer: The differences $D_i = X_{1,i} - X_{2,i}$ are:

Engineer	1	2	3	4	5
Before $X_{1,i}$	43	82	77	39	51
After $X_{2,i}$	51	84	74	48	53
Difference D_i	-8	-2	3	-9	-2
Engineer	6	7	8	9	10
Before $X_{1,i}$	66	55	61	79	43
After $X_{2,i}$	61	59	75	82	48
Difference D_i	5	-4	-14	-3	-5

The observed sample mean is $\overline{d} = -3.9$, and the observed sample variance is $s_D^2 = 31.21$.

The test statistic is:

$$T_0 = \frac{\overline{D} - 0}{S_D/\sqrt{n}} \sim t(n-1),$$

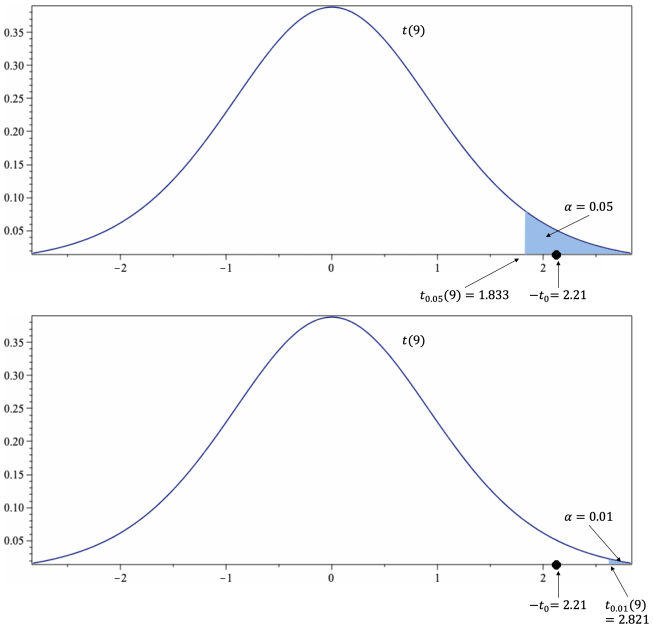
with observed value:

$$t_0 = \frac{-3.9}{\sqrt{31.21/10}} \approx -2.21.$$

We compute

$$P(\overline{D} \leq -3.9) = P(T(9) \leq -2.21) = P(T(9) > 2.21).$$

But $t_{0.05}(9) = 1.833 < t_0 = 2.21 < t_{0.01}(9) = 2.821$, so we reject H_0 at $\alpha = 0.05$, but not at $\alpha = 0.01$.



Unpaired Test Let $X_{1,1}, \ldots, X_{1,n}$ be a random sample from a normal population with unknown mean μ_1 and variance σ_1^2; let $Y_{2,1}, \ldots, Y_{2,m}$ be a random sample from a normal population with unknown mean μ_2 and variance σ_2^2, with both populations **independent** of one another.

We want to test for

$$H_0 : \mu_1 = \mu_2 \text{ against } H_1 : \mu_1 \neq \mu_2.$$

Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$. As always, the observed values are denoted by lower case letters: $\overline{x}$, $\overline{y}$.

σ_1^2 and σ_2^2 Known We can follow the same steps as for the earlier test, with some modifications:

Alternative Hypothesis	Critical Region
$H_1 : \mu_1 > \mu_2$	$z_0 > z_\alpha$
$H_1 : \mu_1 < \mu_2$	$z_0 < -z_\alpha$
$H_1 : \mu_1 \neq \mu_2$	$\lvert z_0 \rvert > z_{\alpha/2}$

where

$$z_0 = \frac{\overline{x} - \overline{y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}},$$

and z_α satisfies $P(Z > z_\alpha) = \alpha$, for $Z \sim \mathcal{N}(0,1)$.

Alt Hypothesis	$p-$Value
$H_1 : \mu_1 > \mu_2$	$P(Z > z_0)$
$H_1 : \mu_1 < \mu_2$	$P(Z < z_0)$
$H_1 : \mu_1 \neq \mu_2$	$2 \cdot \min\{P(Z > z_0), P(Z < z_0)\}$

Example: a sample of $n = 100$ Albertans yields a sample mean income of $\overline{X} = 33,000$\$. A sample of $m = 80$ Ontarians yields $\overline{Y} = 32,000$\$. From previous studies, it is known that the population income standard deviations are, respectively, $\sigma_1 = 5000$\$ in Alberta and $\sigma_2 = 2000$\$ in Ontario. Do Albertans earn more than Ontarians, on average?

Answer: we test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$. The observed difference is $\overline{X} - \overline{Y} = 1000$; the observed test statistic is

$$z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} = \frac{1000}{\sqrt{5000^2/100 + 2000^2/80}} = 1.82;$$

the corresponding p—value is

$$P\left(\overline{X} - \overline{Y} > 1000\right) = P(Z > 1.82) = 0.035,$$

and so we reject H_0 when $\alpha = 0.05$, but not when $\alpha = 0.01$.

σ_1^2 and σ_2^2 Unknown, with Small Samples In this case, the modifications are:

Alternative Hypothesis	Critical Region
$H_1 : \mu_1 > \mu_2$	$t_0 > t_\alpha(n+m-2)$
$H_1 : \mu_1 < \mu_2$	$t_0 < -t_\alpha(n+m-2)$
$H_1 : \mu_1 \neq \mu_2$	$\lvert t_0 \rvert > t_{\alpha/2}(n+m-2)$

where

$$t_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2/n + S_p^2/m}} \quad \text{and} \quad S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2},$$

$t_\alpha(n+m-2)$ satisfies $P(T > t_\alpha(n+m-2)) = \alpha$, and $T \sim t(n+m-2)$.

Alt Hypothesis	p—Value
$H_1 : \mu_1 > \mu_2$	$P(T > t_0)$
$H_1 : \mu_1 < \mu_2$	$P(T < t_0)$
$H_1 : \mu_1 \neq \mu_2$	$2 \cdot \min\{P(T > t_0), P(T < t_0)\}$

Example: a researcher wants to test whether, on average, a new fertilizer yields taller plants. Plants were divided into two groups: a control group treated with an old fertilizer and a study group treated with the new fertilizer. The following data are obtained:

Sample Size	Sample Mean	Sample Variance
$n = 8$	$\overline{X} = 43.14$	$S_1^2 = 71.65$
$m = 8$	$\overline{Y} = 47.79$	$S_2^2 = 52.66$

Test for $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 < \mu_2$.

Answer: the observed difference is $\overline{X} - \overline{Y} = -4.65$ and the **pooled sampled variance** is

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2} = \frac{7(71.65) + 7(52.66)}{8+8-2} = 62.155 = 7.88^2.$$

The observed test statistic is

$$t_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2/n + S_p^2/m}} = \frac{-4.65}{7.88\sqrt{1/8 + 1/8}} = -1.18;$$

the corresponding p—value is

$$P\left(\overline{X} - \overline{Y} < -4.65\right) = P(T(14) < -1.18) = P(T(14) > 1.18) \in (0.1, 0.25)$$

(according to the table), and we do not reject H_0 when $\alpha = 0.05$, or when $\alpha = 0.01$.

σ_1^2 and σ_2^2 Unknown, with Large Samples In this case, the modifications are:

Alternative Hypothesis	Critical Region
$H_1 : \mu_1 > \mu_2$	$z_0 > z_\alpha$
$H_1 : \mu_1 < \mu_2$	$z_0 < -z_\alpha$
$H_1 : \mu_1 \neq \mu_2$	$\lvert z_0 \rvert > z_{\alpha/2}$

where

$$z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n + S_2^2/m}},$$

and z_α satisfies $P(Z > z_\alpha) = \alpha$, for $Z \sim \mathcal{N}(0,1)$.

Alt Hypothesis	p—Value
$H_1 : \mu_1 > \mu_2$	$P(Z > z_0)$
$H_1 : \mu_1 < \mu_2$	$P(Z < z_0)$
$H_1 : \mu_1 \neq \mu_2$	$2 \cdot \min\{P(Z > z_0), P(Z < z_0)\}$

Example: consider the same set-up as in the previous example, but with larger sample sizes: $n = m = 100$. Now test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$.

Answer: the observed difference is (still) -4.65. The observed test statistic is

$$z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n + S_2^2/m}} = \frac{-4.65}{\sqrt{71.65^2/100 + 52.66^2/100}} = -4.17;$$

the corresponding p—value is

$$P\left(\overline{X} - \overline{Y} < -4.65\right) = P(Z < -4.17) \approx 0.0000;$$

and we reject H_0 when either $\alpha = 0.05$ or $\alpha = 0.01$.

8.6 Difference of Two Proportions

As always, we can transfer these tests to proportions, using the normal approximation to the binomial distribution.

For instance, to test for $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$ in samples of size n_1, n_2, respectively, we use the **observed sample difference of proportions**

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{1/n_1 + 1/n_2}},$$

where $\hat{p}$ is the **pooled proportion**

$$\hat{p} = \frac{n_1}{n_1+n_2}\hat{p}_1 + \frac{n_2}{n_1+n_2}\hat{p}_2.$$

and the p—value $2 \cdot \min\{P(Z > z_0), P(Z < z_0)\}$.

9. Miscellanea

Introductory statistical analysis courses usually end with hypothesis testing. In more advanced courses, learners may be introduced to:

- regression analysis and its various extensions [13,19];
- design of experiments and analysis of variance/covariance [13];
- survey sampling methods;
- Bayesian analysis [3,20];
- time series analysis and control charts;
- categorical analysis;
- multivariate analysis [11,13];
- nonparametric methods [5,8];
- advanced probability models [17];
- measure theory, etc.

In this section, we will provide a brief introduction to **simple linear regression** and the **analysis of variance**. The various concepts will be illustrated via the fuels dataset introduced in Section 5.3.

9.1 Linear Regression

Regression analysis can be used to describe the relationship between a **predictor variable** (or regressor) X and a **response variable** Y.

We assume that they are indeed related through the linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where ε is a **random error** and β_0, β_1 are the **regression coefficients**. It is further assumed that $\mathrm{E}[\varepsilon] = 0$, and that the error's variance $\sigma_\varepsilon^2 = \sigma^2$ stays constant when x varies.

The regression model can then be re-written as

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X.$$

Suppose that we have observations $(x_i, y_i),\ i = 1, \ldots, n$, so that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

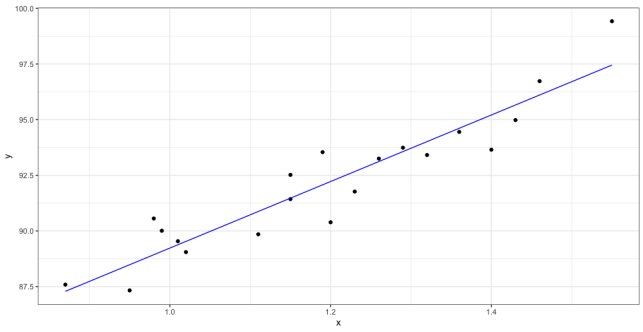
The aim of regression analysis is to find optimal **estimators** b_0, b_1 of the unknown parameters β_0, β_1, in order to obtain the **estimated (fitted) least squares regression line**

$$\hat{y}_i = b_0 + b_1 x_i.$$

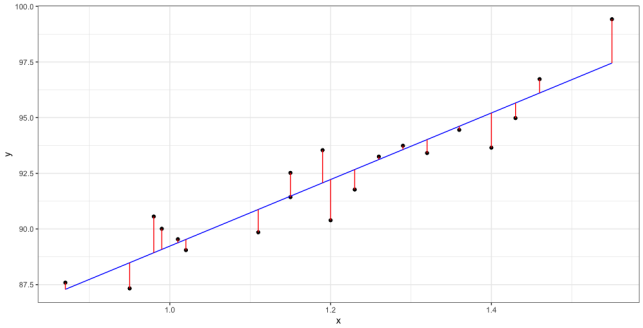
The **residual** or error in predicting y_i using $\hat{y}_i$ is thus

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i, \quad i = 1, \ldots, n.$$

How do we find the estimators? How do we determine if the fitted line is a good model for the data? In the fuels example, for instance, the regression line is $\hat{y} = 74.28 + 14.95x$:



and the residuals $e_i = y_i - \hat{y}_i$ are shown in red below:



Consider the **Sum of Squared Errors** (SSE):

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$[35]

The optimal values of b_0 and b_1 are those that **minimize the SSE**: after all, if the sum of the squared residuals is small, so would be the sum of the residuals.

As such, solving

$$\begin{aligned} 0 = \frac{\mathrm{dSSE}}{\mathrm{d}b_0} &= -2\sum(y_i - b_0 - b_1 x_i) \\ &= -2n(\overline{y} - b_0 - b_1 \overline{x}) \\ 0 = \frac{\mathrm{dSSE}}{\mathrm{d}b_1} &= -2\sum(y_i - b_0 - b_1 x_i) x_i \\ &= -2\Big(\sum x_i y_i - n b_0 \overline{x} - b_1 \sum x_i^2\Big) \end{aligned}$$

yields the **least squares estimators** b_0, b_1 of β_0, β_1.[36]

From $\frac{\mathrm{dSSE}}{\mathrm{d}b_0} = 0$, we see that

$$\overline{y} - b_0 - b_1 \overline{x} = 0 \implies b_0 = \overline{y} - b_1 \overline{x}.$$

For the second coefficient, note that

$$\begin{aligned} S_{xy} &= \sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - n\overline{x}\,\overline{y} \quad \text{and} \\ S_{xx} &= \sum(x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2 \end{aligned}$$

[35] It can be shown that $\mathrm{SSE}/\sigma^2 \sim \chi^2(n-2)$, but that's out-of-scope for this document.

[36] We are **minimizing** the sum of **squared** residuals, hence "least squares".

can be re-written as

$$\sum x_i y_i = S_{xy} + n\overline{x}\,\overline{y} \quad \text{and} \quad \sum x_i^2 = S_{xx} + n\overline{x}^2.$$

From $\frac{\mathrm{dSSE}}{\mathrm{d}b_1} = 0$, we can thus see that

$$\begin{aligned}
&\sum x_i y_i - n b_0 \overline{x} - b_1 \sum x_i^2 = 0 \\
&\therefore (S_{xy} + n\overline{x}\,\overline{y}) - n b_0 \overline{x} - b_1 (S_{xx} + n\overline{x}^2) = 0 \\
&\therefore S_{xy} + n\overline{x}\,\overline{y} - n(\overline{y} - b_1\overline{x})\overline{x} - b_1 S_{xx} - n b_1 \overline{x}^2 = 0 \\
&\therefore S_{xy} + n\overline{x}\,\overline{y} - n\overline{x}\,\overline{y} + n b_1 \overline{x}^2 - b_1 S_{xx} - n b_1 \overline{x}^2 = 0 \\
&\therefore S_{xy} - b_1 S_{xx} = 0 \implies b_1 = \frac{S_{xy}}{S_{xx}}.
\end{aligned}$$

We can also show that the estimators are linear combinations of the observed responses y_i, a fact which can be useful in some of the more advanced proofs:

$$b_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} u_i y_i, \quad b_0 = \overline{y} - b_1 \overline{x} = \sum_{i=1}^{n} v_i y_i.$$

Example: for the fuels data, we already know that

$$S_{xy} \approx 10.18,\ S_{xx} \approx 0.68, \text{and } S_{yy} = 173.38$$

(see Section 5.3). Thus, $b_1 = \frac{10.18}{0.68} = 14.95$. Since

$$n = 20,\ \overline{x} = 1.20, \text{and } \overline{y} = 92.16,$$

we get $b_0 = 92.16 - 20(1.20) = 74.28$. Consequently, the **fitted regression line** is

$$\hat{y} = 74.28 + 14.95x,$$

as claimed on the previous page.

Estimating σ^2 Recall that, by assumption, the variance of the error term is $\sigma_\varepsilon^2 = \sigma^2$. By definition,

$$\mathrm{Var}[\varepsilon] = \mathrm{E}[\varepsilon^2] - \underbrace{\mathrm{E}^2[\varepsilon]}_{=0^2=0} = \mathrm{E}[\varepsilon^2].$$

The best estimator available for the variance must be some average of the squared residuals, of the form

$$\frac{\mathrm{SSE}}{\aleph} = \frac{1}{\aleph}\sum_{i=1}^{n} e_i^2 = \frac{1}{\aleph}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

For a population, the denominator is $\aleph = n$; for a sample, it is $\aleph = n-1$. For the regression error, the **unbiased estimator** of σ^2 is in fact

$$\hat{\sigma}^2 = \mathrm{MSE} = \frac{\mathrm{SSE}}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2},$$

where the SSE has $n-2$ **degrees of freedom**.[37]

[37] The -2 appears because 2 parameters had to be estimated in order to obtain $\hat{y}_i$: b_0 and b_1. In contrast, the sample variance has the denominator is $n-1$ because the data has to be first used to estimate one parameter, the sample mean.

Example: what is the estimated variance of the noise in the linear model for the fuels data?

Answer: $S_{xy} = 10.18,\ S_{yy} = 173.38,\ b_1 = 14.95,\ n = 20$, so

$$\hat{\sigma}^2 = \frac{173.38 - 14.95(10.18)}{20-2} \approx 1.18.$$

Properties of the Least Square Estimators Recall that the simple linear regression model is

$$Y = \beta_0 + \beta_1 X + \varepsilon, \text{ with } \mathrm{E}[\varepsilon] = 0, \sigma_\varepsilon^2 = \sigma^2.$$

Given X, Y is a random variable with mean $\beta_0 + \beta_1 X$ and variance σ^2:

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X, \quad \mathrm{Var}[Y|X] = \sigma^2.$$

Note that b_0 and b_1 depend on the observed x's and y's, which are realizations of the random variables X and Y.

As a result, the **estimators are random variables**, that is to say: different realizations (observed data) lead to different estimates b_0, b_1 for β_0, β_1.

Since b_0, b_1 are linear functions of the observed (independent) responses y_i. it can be shown that

$$\begin{aligned}
&\mathrm{E}[b_0] = \beta_0, \qquad \sigma_{b_0}^2 = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right] = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n S_{xx}}, \\
&\mathrm{E}[b_1] = \beta_1, \qquad \sigma_{b_1}^2 = \sigma^2 / S_{xx}.
\end{aligned}$$

We say that b_0, b_1 are **unbiased estimators** of β_0, β_1. The **estimated standard errors** are obtained by replacing σ^2 by $\mathrm{MSE} = \hat{\sigma}^2$ in the expressions for $\sigma_{b_1}^2$ and $\sigma_{b_0}^2$ above:

$$\mathrm{se}(b_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]} \quad \text{and} \quad \mathrm{se}(b_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

Example: find the estimated standard error for b_0 and b_1 in the fuels data.

Answer: we have $n = 20$, $\overline{x} = 1.20$, $S_{xx} = 0.68$, and $\hat{\sigma}^2 = 1.18$, so that

$$\begin{aligned}
&\mathrm{se}(b_0) = \sqrt{1.18\left[\frac{1}{20} + \frac{1.20^2}{0.68}\right]} \approx 1.593 \quad \text{and} \\
&\mathrm{se}(b_1) = \sqrt{\frac{1.18}{0.68}} \approx 1.317.
\end{aligned}$$

Hypothesis Testing for Linear Regression Armed with standard errors, we can now **test hypotheses** on the regression parameters or the regression as a whole. We can try to ascertain if the evidence supports certain conclusions:

- do the true parameters β_0, β_1 take on specific values;
- does the line of best fit describe the dataset well;
- etc.

The steps are the same as those in Section 8:

1. set up a null hypothesis H_0 and an alternative hypothesis H_1;
2. chose a significance level α;
3. compute the observed value of a specific test statistic (often *via* some form of standardizing);
4. find the critical region or the $p-$value for the test statistic under H_0;
5. reject or fail to reject H_0 based on the critical region or the $p-$value.

Hypothesis Test for the Intercept β_0 We might be interested in testing whether the true intercept β_0 is equal to some **candidate value** $\beta_{0,0}$, i.e. testing for

$$H_0 : \beta_0 = \beta_{0,0} \text{ against } H_1 : \beta_0 \neq \beta_{0,0}.$$

In order to do so, the linear regression model requires **normal errors**

$$\varepsilon \sim \mathcal{N}(0, \sigma^2),$$

which implies that

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2), \quad i = 1, \ldots, n.$$

Since b_0 is a linear function of the observed normal responses y_i, it has itself a normal distribution

$$b_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2}{nS_x x}\sum x_i^2\right).$$

Therefore, under H_0,

$$Z_0 = \frac{b_0 - \beta_{0,0}}{\sqrt{\sigma^2 \frac{\sum x_i^2}{nS_{xx}}}} \sim \mathcal{N}(0,1).$$

But σ^2 is not known; the test statistic obtained by using $\hat{\sigma}^2 = \text{MSE}$ instead of σ^2 is

$$T_0 = \frac{b_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \frac{\sum x_i^2}{nS_{xx}}}} \sim t(n-2),$$

which now follows a Student $t-$distribution with $n-2$ degrees of freedom.[38]

The critical region in this case is provided by one of:

Alternative Hypothesis	Critical/Rejection Region
$H_1 : \beta_0 > \beta_{0,0}$	$t_0 > t_\alpha(n-2)$
$H_1 : \beta_0 < \beta_{0,0}$	$t_0 < -t_\alpha(n-2)$
$H_1 : \beta_0 \neq \beta_{0,0}$	$\lvert t_0\rvert > t_{\alpha/2}(n-2)$

where t_0 is the observed value of T_0 and $t_\alpha(n-2)$ is the $t-$value satisfying $P(T > t_\alpha(n-2)) = \alpha$, for $T \sim t(n-2)$.

As always, we **reject H_0 if t_0 in the critical region**, and **opt not to reject H_0 otherwise**.

[38] Less is known in that case, hence the need to introduce the degrees of freedom as an additional parameter.

Hypothesis Test for the Slope β_1 We might also be interested in testing whether the true slope β_1 is equal to some **candidate value** $\beta_{1,0}$, i.e.

$$H_0 : \beta_1 = \beta_{1,0} \text{ against } H_1 : \beta_1 \neq \beta_{1,0}.$$

The same assumption of normal errors is required, leading to

$$b_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

Therefore, under H_0,

$$Z_0 = \frac{b_1 - \beta_{1,0}}{\sqrt{\sigma^2/S_{xx}}} \sim \mathcal{N}(0,1).$$

But σ^2 is not known; the test statistic obtained by using $\hat{\sigma}^2 = \text{MSE}$ instead of σ^2 is

$$T_0 = \frac{b_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t(n-2),$$

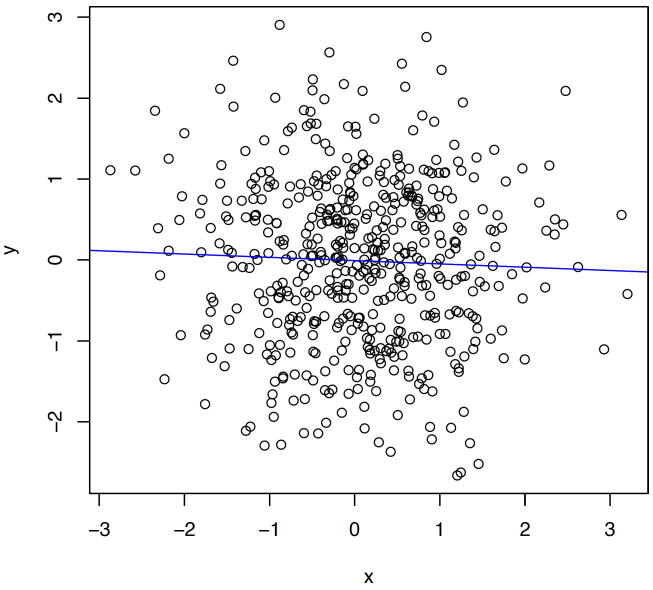
which also follows a Student $t-$distribution with $n-2$ d.f. The critical region in this case is provided by one of:

Alternative Hypothesis	Critical/Rejection Region
$H_1 : \beta_1 > \beta_{1,0}$	$t_0 > t_\alpha(n-2)$
$H_1 : \beta_1 < \beta_{1,0}$	$t_0 < -t_\alpha(n-2)$
$H_1 : \beta_1 \neq \beta_{1,0}$	$\lvert t_0\rvert > t_{\alpha/2}(n-2)$

where t_0 and $t_\alpha(n-2)$ are as in the previous column. The decision rule is identical: we **reject H_0 if t_0 in the critical region**, and **opt not to reject H_0 otherwise**.

Significance of Regression As long as $S_{xx} \neq 0$ (which is to say, as long as there are at least two distinct values of X in the data), we can fit a regression line to the observations using the **least squares framework**.

One of the goals of linear regression is to **describe the linear relationship** between two variables X and Y... assuming that one indeed exists. How can this be done?



The regression line for the dataset on the previous page is

$$\hat{y} = -0.01 - 0.04x.$$

The regression line **exists**, but it does not describe the bivariate data set at all. The relationship between X and Y in that dataset is simply not linear.[39]

Given a regression line, we may want to test whether it is **significant**. The test for **significance of the regression** is

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0.$$

If we reject H_0 in favour of H_1, then the evidence suggests that there is at least a partly linear relationship between X and Y.

Example: in the fuels dataset, we have $b_1 = 14.95$, $n = 20$, $S_{xx} = 0.68$, $\hat{\sigma}^2 = 1.18$. We test for significance of the regression at $\alpha = 0.01$:

$$H_0 : \beta_1 = 0\,, \text{ against } H_1 : \beta_1 \neq 0.$$

Since the observed value of the test statistic is

$$t_0 = \frac{b_1 - 0}{\sqrt{\hat{\sigma}^2/S_{xx}}} = 11.35 > 2.88 = t_{0.01/2}(18)\,,$$

where $t_{0.01/2}(18)$ is the critical value of the $t-$distribution with 18 degrees of freedom at $\alpha = 0.01$ for two-sided tests,[40] we reject H_0 and conclude that there is indeed a linear relationship between X and Y (at $\alpha = 0.01$).[41]

Confidence and Prediction Intervals for Linear Regression We can also build **confidence intervals** (C.I.) for the regression parameters and **prediction intervals** (P.I.) for the predicted values, with the same steps as in Section 7:

1. compute a point estimate W for the parameter β or the prediction Y using the observed data;
2. find the appropriate standard error $\text{se}(W)$;
3. select a confidence level α and find the appropriate critical value $k_{\alpha/2}$, where k represents the corresponding distribution, and
4. build the $100(1-\alpha)\%$ interval $W \pm k_{\alpha/2} \cdot \text{se}(W)$.

C.I. for the Intercept β_0 and the Slope β_1 Since we estimate the error variance with $\hat{\sigma}^2 = \text{MSE}$, we need to use Student's $t-$distribution with $n-2$ degrees of freedom (remember that we use the data to estimate 2 parameters).

The $100(1-\alpha)\%$ C.I. for β_0 and β_1 are:

$$\begin{aligned}\beta_0 : \quad & b_0 \pm t_{\alpha/2}(n-2) \cdot \text{se}(b_0) \\ & = b_0 \pm t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2 \frac{\sum x_i^2}{n S_{xx}}}\end{aligned}$$

and

$$\begin{aligned}\beta_1 : \quad & b_1 \pm t_{\alpha/2}(n-2) \cdot \text{se}(b_1) \\ & = b_1 \pm t_{\alpha/2}(n-2)\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\end{aligned}$$

The caveat regarding the interpretation of confidence intervals still applies.

Example: build 95%, 99% C.I. for β_0, β_1 in the fuels data.

Answer: $b_0 = 74.283$, $b_1 = 14.947$, $\text{se}(b_0) = 1.593$, $\text{se}(b_1) = 1.317$, $t_{0.025}(18) = 2.10$ and $t_{0.005}(18) = 2.88$, as we have seen in previous examples.

Then, for $\alpha = 0.05$, we have

$$\begin{aligned}\beta_0 : \ & 74.283 \pm 2.10(1.593) = (70.93, 77.63) \\ \beta_1 : \ & 14.497 \pm 2.10(1.317) = (12.18, 17.71)\end{aligned}$$

and for $\alpha = 0.01$, we have

$$\begin{aligned}\beta_0 : \ & 74.283 \pm 2.88(1.593) = (69.70, 78.87) \\ \beta_1 : \ & 14.497 \pm 2.88(1.317) = (11.15, 18.74).\end{aligned}$$

Confidence Intervals for the Mean Response We might also be interested in estimating $\mu_{Y|x_0} = \text{E}[Y|x_0]$, the **mean response** at an observed x_0 (in practice, there could be more than one response at the predictor, due to replication in an experiment, say).

The predicted value can be read directly from the regression line:

$$\hat{\mu}_{Y|x_0} = b_0 + b_1 x_0.$$

The distance (at x_0) between the estimated value and the true regression line is

$$\hat{\mu}_{Y|x_0} - \mu_{Y|x_0} = (b_0 - \beta_0) + (b_1 - \beta_1)\, x_0.$$

The predicted value $\hat{\mu}_{Y|x_0}$ will depend on the observed values, and so it has a distribution. We can show that $\text{E}[\hat{\mu}_{Y|x_0}] = \mu_{Y|x_0}$ and

$$\text{Var}[\hat{\mu}_{Y|x_0}] = \text{Var}[b_0 + b_1 x_0] = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right].$$

Note that $\text{Var}[b_0 + b_1 x_0] \neq \text{Var}[b_0] + \text{Var}[b_1 x_0]$ since b_0 and b_1 are dependent.

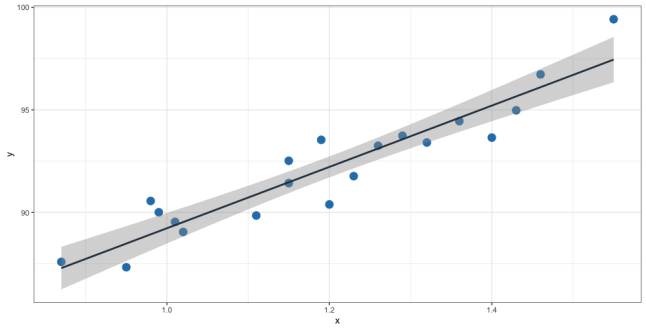
With the usual $t_{\alpha/2}(n-2)$, the $100(1-\alpha)\%$ C.I. for the **mean response** $\mu_{Y|x_0}$ (or for the line of regression) is

$$\hat{\mu}_{Y|x_0} \pm t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}.$$

Example: for the fuels dataset, the 95% C.I. for $\mu_{Y|x_0}$ is

$$74.28 + 14.95 x_0 \pm 2.10\sqrt{1.18\left[\frac{1}{20} + \frac{(x_0 - 1.12)^2}{0.68}\right]}.$$

[39] It is more like a "blob."
[40] Obtained with `-qt (0.01/2,18)` in R.
[41] As is readily apparent in the scatterplot on p. 52.



A fair number of the observations are found outside the 95% C.I. for the mean response, potentially because of the relatively small sample size.

Predicting New Observations If x_0 is the value of interest for the regressor (predictor), then the estimated value of the response variable Y is

$$\hat{y} = \hat{Y}_0 = b_0 + b_1 x_0.$$

If Y_0 is the **true future observation** at $X = x_0$ (so, if $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon$) and $\hat{Y}_0$ is the predicted value, given by the above equation, then we can show that the prediction error

$$e_{\hat{p}} = Y_0 - \hat{Y}_0 = (\beta_0 - b_0) + (\beta_1 - b_1)x_0 + \varepsilon$$

follows a normal distribution with zero mean and variance

$$\sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right].$$

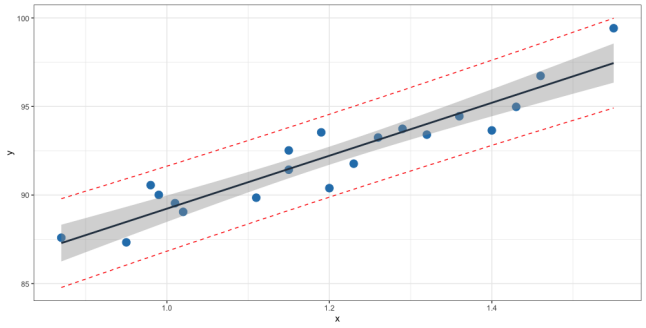
If we substitute σ^2 by its estimator $\hat{\sigma}^2 = \text{MSE}$, we get a $100(1-\alpha)\%$ **prediction interval** for Y_0:

$$b_0 + b_1 x_0 \pm t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right]},$$

where $t_{\alpha/2}$ is the critical value of Student's $t-$distribution with $n-2$ degrees of freedom at confidence level α.

Example: for the fuels dataset, the 95% P.I. for $\mu_{Y|x_0}$ is

$$74.28 + 14.95 x_0 \pm 2.10\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(x_0 - 1.12)^2}{0.68} \right]}.$$



None of the observations are found outside the 95% P.I. for new observations. In general, for a given α, the prediction interval is wider than the confidence interval, which should not surprising: the CLT implies that the mean response has a smaller variance than the predicted responses.

9.2 Analysis of Variance

The test for **significance of regression**,

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0,$$

can be restated in term of the **analysis-of-variance** (ANOVA), provided by the following table:

Source of Variation	Sum of Squares	df	Mean Square	F^*	$p-$Value
Regression	SSR	1	MSR	$\frac{\text{MSR}}{\text{MSE}}$	$P(F > F^*)$
Error	SSE	$n-2$	MSE		
Total	SST	$n-1$			

In this table, the $F-$statistic $F^* \sim F(1, n-2)$, and

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad \text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2, \quad \text{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2,$$

$$\text{MSR} = \frac{\text{SSR}}{1}, \quad \text{MSE} = \frac{\text{SSE}}{n-2}, \quad \text{and} \quad F^* = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/1}{\text{SSE}/n-2}$$

The **rejection region** for the null hypothesis $H_0 : \beta_1 = 0$ is

$$\left| \frac{b_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \right| > t_{\alpha/2}(n-2),$$

but it can also be written as $F^* > f_\alpha(1, n-2)$, where $f_\alpha(1, n-2)$ is the critical $F-$value of the F-distribution with $\nu_1 = 1$ and $\nu_2 = n-2$ df.

Example: the $F-$statistic for the ANOVA of the fuels data set can be computed directly from the data: $F^* = 128.9$. The numbers of df are $\nu_1 = 1$ and $\nu_2 = 20 - 2 = 18$.

The critical value at $\alpha = 0.05$ is $f_{0.05}(1, 18) = 4.41$.[42] Since $F^* = 128.9 > f_{0.05}(1, 18) = 4.4$, we reject the null hypothesis H_0 in favour of the regression being significant at $\alpha = 0.05$.

Coefficient of Determination The **coefficient of determination** is the expression

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}},$$

where SSE and SST are as in the ANOVA table.

It is the proportion of the variability in the response that is explained by the fitted model. Note that R^2 always lies between 0 and 1; when $R^2 \approx 1$, the fit is considered to be very good.[43]

[42] Obtained with `=qf(0.95,1,18)`.

[43] **BE CAREFUL**: in practice, R^2 is not always the best way to determine the **goodness-of-fit** of the regression. There are other factors (such as the number of observations) which can affect the coefficient of determination.

10. Exercises

Throughout, “NOTP” means "none of the preceding".

1. Two events each have probability 0.2 of occurring and are independent. The probability that neither occur is:

   a) 0.64 b) 0.04 c) 0.2 d) 0.4 e) NOTP

2. Two events each have probability 0.2 and are mutually exclusive. The probability that neither occurs is:

   a) 0.36 b) 0.04 c) 0.2 d) 0.6 e) NOTP

3. A smoke-detector system consists of two parts A and B. If smoke occurs then the item A detects it with probability 0.95, the item B detects it with probability 0.98 whereas both of them detect it with probability 0.94. What is the probability that the smoke will not be detected?

   a) 0.01 b) 0.99 c) 0.04 d) 0.96 e) NOTP

4. Three football players will attempt to kick a field goal. Let A_1, A_2, A_3 denote the events that the field goal is made by player $1, 2, 3$, respectively. Assume that A_1, A_2, A_3 are independent and $P(A_1) = 0.5$, $P(A_2) = 0.7$, $P(A_3) = 0.6$. Compute the probability that exactly 1 player is successful.

   a) 0.29 b) 0.21 c) 0.71 d) 0.79 e) NOTP

5. In a group of 16 candidates, 7 are chemists and 9 are physicists. In how many ways can one choose a group of 5 candidates with 2 chemists and 3 physicists?

6. There is a theorem of combinatorics that states that the number of permutations of n objects in which n_1 are alike of kind 1, n_2 are alike of kind 2, ..., and n_r are alike of kind r (that is, $n = n_1 + n_2 + \cdots + n_r$) is

   $$\frac{n!}{n_1! \cdot n_2! \cdot \cdots \cdot n_r!}.$$

   Find the number of different words that can be formed by rearranging the letters in the following words.

   a) NORMAL b) HHTTTT c) ILLINI d) MISSISSIPPI

7. A class consists of 490 engineering and 510 science students. The students are divided according to their marks:

	Passed	Failed
Eng.	430	60
Sci.	410	100

   If one person is selected randomly, what is the probability that they failed if they were an engineering student?

   a) 0.06 b) 0.12 c) 0.41 d) 0.81 e) NOTP

8. A company which produces a particular drug has two factories, A and B. 70% of the drug are made in factory A, 30% in factory B. Suppose that 95% of the drugs produced by factory A meet standards while only 75% of those produced by factory B meet standards. What is the probability that a random dose meets standards?

   a) 0.81 b) 0.95 c) 0.75 d) 0.7 e) NOTP

9. A medical research team wished to evaluate a proposed screening test for Alzheimer’s disease. The test was given to a random sample of 450 patients with Alzheimer’s disease; in 436 cases the test result was positive. The test was also given to a random sample of 500 patients without the disease; only in 5 cases was the result was positive. It is known that in Canada 11.3% of the population aged 65+ have Alzheimer’s disease. Find the probability that a person has the disease given that their test was positive (choose the closest answer).

   a) 0.97 b) 0.93 c) 0.99 d) 0.07 e) NOTP

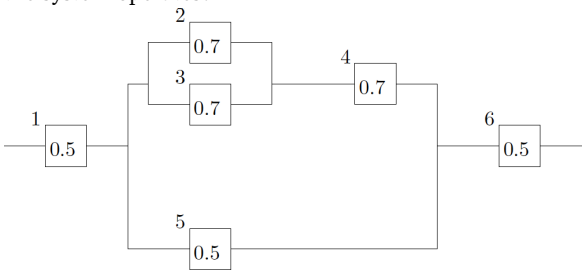
10. Twelve items are independently sampled from a production line. If the probability that any given item is defective is 0.1, the probability of at most two defectives in the sample is closest to ...

    a) 0.39 b) 0.99 c) 0.74 d) 0.89 e) NOTP

11. A student can solve 6 problems from a list of 10. For an exam 8 questions are selected at random from the list. What is the probability that the student will solve exactly 5 problems?

    a) 0.98 b) 0.02 c) 0.28 d) 0.53 e) NOTP

12. Consider the following system with six components. We say that it is functional if there exists a path of functional components from left to right. The probability of each component functions is shown. Assume that the components function or fail independently. What is the probability that the system operates?



    a) 0.18 b) 0.82 c) 0.64 d) 0.20 e) NOTP

13. Pieces of aluminum are classified according to the finishing of the surface and according to the finishing of edge. The results from 85 samples are summarized as follows:

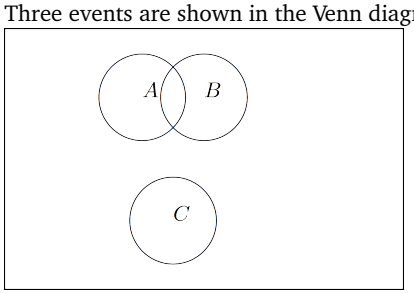
	Edge	
Surface	excellent	good
excellent	60	5
good	16	4

    Let A denote the event that a selected piece has "excellent" surface, and let B denote the event that a selected piece has “excellent” edge. If samples are elected randomly, determine the following probabilities:

    a) $P(A)$ b) $P(B)$ c) $P(A^c)$

    d) $P(A \cap B)$ e) $P(A \cup B)$ f) $P(A^c \cup B)$

14. Three events are shown in the Venn diagram below.



Shade the region corresponding to the following events:

a) A^c b) $(A \cap B) \cup (A \cap B^c)$

c) $(A \cap B) \cup C$ d) $(B \cup C)^c$

e) $(A \cap B)^c \cup C$

15. If $P(A) = 0.1$, $P(B) = 0.3$, $P(C) = 0.3$, and events A, B, C are mutually exclusive, determine the following probabilities:

a) $P(A \cup B \cup C)$ b) $P(A \cap B \cap C)$ c) $P(A \cap B)$

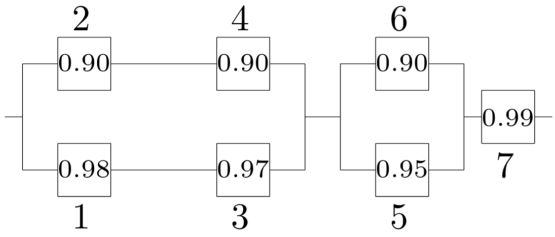
d) $P((A \cup B) \cap C)$ e) $P(A^c \cap B^c \cap C^c)$ f) $P[(A \cup B \cup C)^c]$

16. The probability that an electrical switch, which is kept in dryness, fails during the guarantee period, is 1%. If the switch is humid, the failure probability is 8%. Assume that 90% of switches are kept in dry conditions, whereas remaining 10% are kept in humid conditions.

a) What is the probability that the switch fails during the guarantee period?

b) If the switch failed during the guarantee period, what is the probability that it was kept in humid conditions?

17. The following system operates only if there is a path of functional device from left to the right. The probability that each device functions is as shown. What is the probability that the circuit operates? Assume independence.



18. An inspector working for a manufacturing company has a 95% chance of correctly identifying defective items and 2% chance of incorrectly classifying a good item as defective. The company has evidence that 1% of the items it produces are nonconforming (defective).

(a) What is the probability that an item selected for inspection is classified as defective?

(b) If an item selected at random is classified as non defective, what is the probability that it is indeed good?

19. Consider an ordinary 52-card North American playing deck (4 suits, 13 cards in each suit).

a) How many different 5—card poker hands can be drawn from the deck?

b) How many different 13—card bridge hands can be drawn from the deck?

c) What is the probability of an all-spade 5—card poker hand?

d) What is the probability of a flush (5—cards from the same suit)?

e) What is the probability that a 5—card poker hand contains exactly 3 Kings and 2 Queens?

f) What is the probability that a 5—card poker hand contains exactly 2 Kings, 2 Queens, and 1 Jack?

20. Students on a boat send messages back to shore by arranging seven coloured flags on a vertical flagpole.

a) If they have 4 orange flags and 3 blue flags, how many messages can they send?

b) If they have 7 flags of different colours, how many messages can they send?

c) If they have 3 purple flags, 2 red flags, and 4 yellow flags, how many messages can they send?

21. The Stanley Cup Finals of hockey or the NBA Finals in basketball continue until either the representative team form the Western Conference or from the Eastern Conference wins 4 games. How many different orders are possible ($WWEEEE$ means that the Eastern team won in 6 games) if the series goes

a) 4 games? b) 5 games? c) 6 games? d) 7 games?

22. Consider an ordinary 52-card North American playing deck (4 suits, 13 cards in each suit), from which cards are drawn at random and without replacement, until 3 spades are drawn.

a) What is the probability that there are 2 spades in the first 5 draws?

b) What is the probability that a spade is drawn on the 6th draw given that there were 2 spades in the first 5 draws?

c) What is the probability that 6 cards need to be drawn in order to obtain 3 spades?

d) All the cards are placed back into the deck, and the deck is shuffled. 4 cards are then drawn from. What is the probability of having drawn a spade, a heart, a diamond, and a club, in that order?

23. A student has 5 blue marbles and 4 white marbles in his left pocket, and 4 blue marbles and 5 white marbles in his right pocket. If they transfer one marble at random from their left pocket to his right pocket, what is the probability of them then drawing a blue marble from their right pocket?

24. An insurance company sells a number of different policies; among these, 60% are for cars, 40% are for homes, and 20% are for both. Let A_1, A_2, A_3, A_4 represent people with only a

car policy, only a home policy, both, or neither, respectively. Let B represent the event that a policyholder renews at least one of the car or home policies.

a) Compute $P(A_1)$, $P(A_2)$, $P(A_3)$, and $P(A_4)$.

b) Assume $P(B \mid A_1) = 0.6$, $P(B \mid A_2) = 0.7$, $P(B \mid A_3) = 0.8$. Given that a client selected at random has a car or a home policy, what is the probability that they will renew one of these policies?

25. An urn contains four balls numbered 1 through 4. The balls are selected one at a time, without replacement. A match occurs if ball m is the mth ball selected. Let the event A_i denote a match on the ith draw, $i = 1, 2, 3, 4$.

a) Compute $P(A_i)$, $i = 1, 2, 3, 4$.

b) Compute $P(A_i \cap A_j)$, $i, j = 1, 2, 3, 4$, $i \neq j$.

c) Compute $P(A_i \cap A_j \cap A_k)$, $i, j, k = 1, 2, 3, 4$, $i \neq j, i \neq k, j \neq k$.

d) What is the probability of at least 1 match?

26. The probability that a company's workforce has at least one accident in a given month is $(0.01)k$, where k is the number of days in the month. Assume that the number of accidents is independent from month to month. If the company's year starts on January 1, what is the probability that the first accident occurs in April?

27. A Pap smear is a screening procedure used to detect cervical cancer. Let T^- and T^+ represent the events that the test is negative and positive, respectively, and let C represent the event that the person tested has cancer. The false negative rate for this test when the patient has the cancer is 16%; the false positive test for this test when the patient does not have cancer is 19%. In North America, the rate of incidence for this cancer is roughly 8 out of 100,000 women. Based on these numbers, is a Pap smear an effective procedure? What factors influence your conclusion?

28. Of three different fair dice, one each is given to Elowyn, Llewellyn, and Gwynneth. They each roll it. Let $E =$ {Elowyn rolls a 1 or a 2}, $LL =$ {Llewellyn rolls a 3 or a 4}, and $G =$ {Gwynneth rolls a 5 or a 6} be events.

a) What are the probabilities of each of E, LL, and G occurring?

b) What are the probabilities of any two of E, LL, and G occurring simultaneously?

c) What is the probability of all three of the events occurring simultaneously?

d) What is the probability of at least one of E, LL, or G occurring?

29. Over the course of two baseball seasons, player A obtained 126 hits in 500 at-bats in Season 1, and 90 hits in 300 at-bats in Season 2; player B, on the other hand, obtained 75 hits in 300 at-bats in Season 1, and 145 hits in 500 at-bats in Season 2. A player's batting average is the number of hits they obtain divided by the number of at-bats.

a) Which player has the best batting average in Season 1? In Season 2?

b) Which player has the best batting average over the 2-year period?

c) What is happening here?

30. A stranger comes to you and shows you what appears to be a normal coin, with two distinct sides: Heads (H) and Tails (T). They flip the coin 4 times and record the following sequence of tosses: $HHHH$.

a) What is the probability of obtaining this specific sequence of tosses? What assumptions do you make along the way in order to compute the probability? What is the probability that the next toss will be a T.

b) The stranger offers you a bet: they will toss the coin another time; if the toss is T, they give you 100$, but if it is H, you give them 10$. Would you accept the bet (if you are not morally opposed to gambling)?

c) Now the stranger tosses the coin 60 times and records $60 \times H$ in a row: $H \cdots H$. They offer you the same bet. Do you accept it?

d) What if they offered 1000$ instead? $1,000,000$?

31. An experiment consists in selecting a bowl, and then drawing a ball from that bowl. Bowl B_1 contains two red balls and four white balls; bowl B_2 contains one red ball and two white balls; and bowl B_3 contains five red balls and four white balls. The probabilities for selecting the bowls are not uniform: $P(B_1) = 1/3$, $P(B_2) = 1/6$, and $P(B_3) = 1/2$, respectively.

a) What is the probability of drawing a red ball $P(R)$?

b) If the experiment is conducted and a red ball is drawn, what is the probability that the ball was drawn from bowl B_1? B_2? B_3?

32. Two companies A and B consider making an offer for road construction. Company A submits a proposal. The probability that B submits a proposal is $1/3$. If B does not submit the proposal, the probability that A gets the job is $3/5$. If B submits the proposal, the probability that A gets the job is $1/3$. What is the probability that A will get the job?

a) 0.67 b) 0.51 c) 0.75 d) 0.33 e) NOTP

33. In a box of 50 fuses there are 8 defective ones. We choose 5 fuses randomly (without replacement). What is the probability that all 5 fuses are not defective?

a) 0.40 b) 0.84 c) 0.37 d) 0.43 e) NOTP

34. The sample space of a random experiment is $\{a, b, c, d, e, f\}$ and each outcome is equally likely. A random variable is defined as follows

outcome	a	b	c	d	e	f
X	0	0	1.5	1.5	2	3

Determine the probability mass function of X. Determine the following probabilities:

a) $P(X = 1.5)$ b) $P(0.5 < X 2.7)$

c) $P(X > 3)$ d) $P(0 \leq X < 2)$

e) $P(X = 0 \text{ or } 2)$

35. Determine the mean and the variance of the random variable defined in the previous question.

36. We say that X has **uniform distribution** on a set of values $\{X_1, \ldots, X_k\}$ if

$$P(X = X_i) = \frac{1}{k}, \qquad i = 1, \ldots, k.$$

The thickness measurements of a coating process are **uniformly distributed** with values 0.15, 0.16, 0.17, 0.18, 0.19. Determine the mean and variance of the thickness measurements. Is this result compatible with a uniform distribution?

37. Samples of rejuvenated mitochondria are mutated in 1% of cases. Suppose 15 samples are studied and that they can be considered to be independent (from a mutation standpoint). Determine the following probabilities:

a) no samples are mutated;
b) at most one sample is mutated, and
c) more than half the samples are mutated.

Use the following CDF table for the $\mathscr{B}(n,p)$, with $n = 15$ and $p = 0.99$:

r	0	1	2	3	4	5	6	7
$P(X \leq r)$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
r	8	9	10	11	12	13	14	15
$P(X \leq r)$	0.0000	0.0000	0.0000	0.0000	0.0004	0.0096	0.1399	1.0000

38. Samples of 20 parts from a metal punching process are selected every hour. Typically, 1% of the parts require re-work. Let X denote the number of parts in the sample that require re-work. A process problem is suspected if X exceeds its mean by more than three standard deviations.

a) What is the probability that there is a process problem?
b) If the re-work percentage increases to 4%, what is the probability that X exceeds 1?
c) If the re-work percentage increases to 4%, what is the probability that X exceeds 1 in at least one of the next five sampling hours?

39. In a clinical study, volunteers are tested for a gene that has been found to increase the risk for a particular disease. The probability that the person carries a gene is 0.1.

a) What is the probability that 4 or more people will have to be tested in order to detect 1 person with the gene?
b) How many people are expected to be tested in order to detect 1 person with the gene?
c) How many people are expected to be tested in order to detect 2 people with the gene?

40. The number of failures of a testing instrument from contaminated particles on the product is a Poisson random variable with a mean of 0.02 failure per hour.

a) What is the probability that the instrument does not fail in an 8—hour shift?
b) What is the probability of at least 1 failure in a 24—hour day?

41. Use R to generate a sample from a binomial distribution and from a Poisson distribution (select parameters as you wish). Use R to compute the sample means and sample variances. Compare these values to population means and population variances.

42. A container of 100 light bulbs contains 5 bad bulbs. We draw 10 bulbs without replacement. Find the probability of drawing at least 1 defective bulb.

a) 0.42 b) 0.58 c) 0.1 d) 0.9 e) NOTP

43. Let X be a discrete random variable with range $\{0, 1, 2\}$ and probability mass function (p.m.f.) given by $f(0) = 0.5$, $f(1) = 0.3$, and $f(2) = 0.2$. The expected value and variance of X are, respectively,

a) .7;.6 b) .7;1.1 c) .5;.6 d) .5;1.1 e) NOTP

44. A factory employs several thousand workers, of whom 30% are not from an English-speaking background. If 15 members of the union executive committee were chosen from the workers at random, evaluate the probability that exactly 3 members of the committee are not from an English-speaking background.

a) 0.17 b) 0.83 c) 0.98 d) 0.51 e) NOTP

Use the following CDF table for the $\mathscr{B}(n,p)$, with $n = 15$ and $p = 0.30$ if needed:

r	0	1	2	3	4	5	6	7
$P(X \leq r)$	0.0047	0.0353	0.1268	0.2969	0.5155	0.7216	0.8689	0.9500
r	8	9	10	11	12	13	14	15
$P(X \leq r)$	0.9848	0.9963	0.9993	0.9999	1.0000	1.0000	1.0000	1.0000

45. Assuming the context of the previous questions, what is the probability that a majority of the committee members do not come from an English-speaking background?

46. In a video game, a player is confronted with a series of opponents and has an 80% probability of defeating each one. Success with any opponent (that is, defeating the opponent) is independent of previous encounters. The player continues until defeated. What is the probability that the player encounters at least three opponents?

a) 0.8 b) 0.64 c) 0.5 d) 0.36 e) NOTP

47. Assuming the context of the previous question, how many encounters is the player expected to have?

a) 5 b) 4 c) 8 d) 10 e) NOTP

48. From past experience it is known that 3% of accounts in a large accounting company are in error. The probability that exactly 5 accounts are audited before an account in error is found, is:

a) 0.242 b) 0.011 c) 0.030 d) 0.026 e) NOTP

49. A receptionist receives on average 2 phone calls per minute. Assume that the number of calls can be modeled using a Poisson random variable. What is the probability that he does not receive a call within a 3—minute interval?

a) e^{-2} b) $e^{-1/2}$ c) e^{-6} d) e^{-1} e) NOTP

50. Roll a 4—sided die twice, and let X equal the larger of the two outcomes if they are different and the common value if they are the same. Find the p.m.f. and the c.d.f. of X.

51. Compute the mean and the variance of X as defined in the previous question, as well as $\mathrm{E}[X(5-X)]$.

52. A basketball player is successful in 80% of her (independent) free throw attempts. Let X be the minimum number of attempts in order to succeed 10 times. Find the p.m.f. of X and the probability that $X = 12$.

53. Let X be the minimum number of independent trials (each with probability of success p) that are needed to observe r successes. The p.m.f. of X is

$$f(x) = P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-1}, \quad x = r, r+1, \ldots$$

The mean and variance of X are

$$\mathrm{E}[X] = \frac{r}{p} \quad \text{and} \quad \mathrm{Var}[X] = \frac{r(1-p)}{p^2}.$$

Compute the mean minimum number of independent free throw attempts required to observe 10 successful free throws if the probability of success at the free thrown line is 80%. What about the standard deviation of X?

54. If $n \geq 20$ and $p \leq 0.05$, it can be shown that the binomial distribution with n trials and an independent probability of success p can be approximated by a Poisson distribution with parameter $\lambda = np$:

$$\frac{(np)^x e^{-np}}{x!} \approx \binom{n}{x} p^x (1-p)^{n-x}.$$

A manufacturer of light bulbs knows that 2% of its bulbs are defective. What is the probability that a box of 100 bulbs contains exactly at most 3 defective bulbs? Use the Poisson approximation to estimate the probability.

55. Consider a discrete random variable X which has a uniform distribution over the first positive m integers, i.e.

$$f(x) = P(X = x) = \frac{1}{m}, \quad x = 1, \ldots, m,$$

and $f(x) = 0$ otherwise. Compute the mean and the variance of X. For what values of m is $\mathrm{E}[X] > \mathrm{Var}[X]$?

56. Assume that arrivals of small aircrafts at an airport can be modeled by a Poisson random variable with an average of 1 aircraft per hour.

    a) What is the probability that more than 3 aircrafts arrive within an hour?

    b) Consider 15 consecutive and disjoint 1—hour intervals. What is the probability that in none of these intervals we have more than 3 aircraft arrivals?

    c) What is the probability that exactly 3 aircrafts arrive within 2 hours?

57. In a group of ten students, each student has a probability of 0.7 of passing the exam. What is the probability that exactly 7 of them will pass an exam?

    a) 0.98 b) 0.27 c) 0.05 d) 0.95 e) NOTP

58. A company's warranty states that the probability that a new swimming pool requires some repairs within the 1st year is 20%. What is the probability, that the sixth sold pool is the first one which requires some repairs within the 1st year?

    a) 0.61 b) 0.39 c) 0.93 d) 0.07 e) NOTP

59. Consider the following R output:

```
> pbinom(16,100,0.25)
[1] 0.02111062
> pbinom(30,100,0.25)
[1] 0.8962128
> pbinom(32,100,0.25)
[1] 0.9554037
> pbinom(15,100,0.25)
[1] 0.01108327
> pbinom(17,100,0.25)
[1] 0.03762626
> pbinom(31,100,0.25)
[1] 0.9306511
```

Let $X \sim \mathcal{B}(n, p)$ with $n = 100$ and $p = 0.25$. Using the R output above, calculate $P(16 \leq X \leq 31)$.

    a) 0.92 b) 0.91 c) 0.93 d) 0.94 e) NOTP

60. Consider a random variable X with probability density function (p.d.f.) given by

$$f(x) = \begin{cases} 0 & \text{if } x \leq -1 \\ 0.75(1-x^2) & \text{if } -1 \leq x < 1 \\ 0 & \text{if } x \geq 1 \end{cases}$$

What is the expected value and the standard deviation of X?

    a) 0;3 b) 0;0.44c) 1;0.2 d) 1;3 e) NOTP

61. A random variable X has a cumulative distribution function (c.d.f.)

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x/2 & \text{if } 0 < x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

What is the mean value of X?

    a) 1 b) 2 c) 0 d) 0.5 e) NOTP

62. Let X be a random variable with p.d.f. $f(x) = \frac{3}{2}x^2$ for $-1 \leq x \leq 1$, and $f(x) = 0$ otherwise. Find $P(X^2 \leq 0.25)$.

    a) 0.250 b) 0.125 c) 0.500 d) 0.061 e) NOTP

63. In the inspection of tin plate produced by a continuous electrolytic process, 0.2 imperfections are spotted per minute, on average. Find the probability of spotting at least 2 imperfections in 5 minutes. Assume that we can model the occurrences of imperfections as a Poisson process.

    a) 0.736 b) 0.264 c) 0.632 d) 0.368 e) NOTP

64. If $X \sim \mathcal{N}(0, 4)$, the value of $P(|X| \geq 2.2)$ is (using the normal table):

    a) 0.23 b) 0.84 c) 0.25 d) 0.27 e) 0.73 f) NOTP

65. If $X \sim \mathcal{N}(10,1)$, the value of k such that $P(X \leq k) = 0.701944$ is closest to

    a) 0.59 b) 0.30 c) 0.53 d) 10.53 e) 10.30 f) 10.59

66. The time it takes a supercomputer to perform a task is normally distributed with mean 10 milliseconds and standard deviation 4 milliseconds. What is the probability that it takes more than 18.2 milliseconds to perform the task? (use the normal table or R).

    a) 0.98 b) 0.85 c) 0.02 d) 0.22 e) 0.55 f) NOTP

67. Let X be a random variable. What is the value of b (where b is not a function of X) which minimizes $\mathrm{E}[(X-b)^2]$?

68. The time to reaction to a visual signal follows a normal distribution with mean 0.5 seconds and standard deviation 0.035 seconds.

    a) What is the probability that time to react exceeds 1 second?

    b) What is the probability that time to react is between 0.4 and 0.5 seconds?

    c) What is the time to reaction that is exceeded with probability of 0.9?

69. Refer to the situation described in question 56.

    d) What is the length of the interval such that the probability of having no arrival within this interval is 0.1?

    e) What is the probability that one has to wait at least 3 hours for the arrival of 3 aircrafts?

    f) What is the mean and variance of the waiting time for 3 aircrafts?

70. Assume that X is normally distributed with mean 10 and standard deviation 3. In each case, find the value x such that:

    a) $P(X > x) = 0.5$

    b) $P(X > x) = 0.95$

    c) $P(x < X < 10) = 0.2$

    d) $P(-x < X - 10 < x) = 0.95$

    e) $P(-x < X - 10 < x) = 0.99$

71. Let $X \sim \mathrm{Exp}(\lambda)$ with mean 10. Find $P(X > 30 \mid X > 10)$.

    a) $1-\exp(-2)$ b) $\exp(-2)$ c) $\exp(-3)$

    d) $1/10$ e) $\exp(-200)$ f) NOTP

72. Consider a random variable X with the following probability density function:

$$f(x) = \begin{cases} 0 & \text{if } x \leq -1 \\ \frac{3}{4}(1-x^2) & \text{if } -1 < x < 1 \\ 0 & \text{if } x \geq 1 \end{cases}$$

    The value of $P(X \leq 0.5)$ is

    a) 11/32 b) 27/32 c) 16/32 d) 1

    e) NOTP

73. A receptionist receives on average 2 phone calls per minute. If the number of calls follows a Poisson process, what is the probability that the waiting time for call will be greater than 1 minute?

    a) $e^{-1/15}$ b) $e^{-1/30}$ c) e^{-2} d) e^{-1} e) NOTP

74. A company manufactures hockey pucks. It is known that their weight is normally distributed with mean 1 and standard deviation 0.05. The pucks used by the NHL must weigh between 0.9 and 1.1. What is the probability that a randomly chosen puck can be used by NHL?

    a) 1 b) 0.95 c) 0.46 d) 0.99 e) NOTP

75. Find $\mathrm{Var}[X]$, $\mathrm{Var}[Y]$, and $\mathrm{Cov}(X,Y)$ for the dice example on page 25. Are X and Y independent?

76. Find $\mathrm{Var}[X_1]$, $\mathrm{Var}[X_2]$, and $\mathrm{Cov}(X_1,X_2)$ for the chip example on page 26. Are X_1 and X_2 independent?

77. Find $\mathrm{Var}[X]$, $\mathrm{Var}[Y]$, and $\mathrm{Cov}(X,Y)$ if X and Y have joint p.m.f.

$$f(x,y) = \frac{x+y}{21}, \quad x = 1,2,3, \quad y = 1,2.$$

78. Find $\mathrm{Var}[X]$, $\mathrm{Var}[Y]$, and $\mathrm{Cov}(X,Y)$ if X and Y have joint p.m.f.

$$f(x,y) = \frac{xy^2}{30}, \quad x = 1,2,3, \quad y = 1,2.$$

    Are X and Y independent?

79. Find $\mathrm{Var}[X]$, $\mathrm{Var}[Y]$, and $\mathrm{Cov}(X,Y)$ if X and Y have joint p.m.f.

$$f(x,y) = \frac{xy^2}{13}, \quad (x,y) = (1,1),(1,2),(2,2)$$

    Are X and Y independent?

80. Find $\mathrm{Var}[X]$, $\mathrm{Var}[Y]$, and $\mathrm{Cov}(X,Y)$ if X and Y have joint p.d.f.

$$f(x,y) = \frac{3}{2}x^2(1-|y|), \quad -1 < x < 1, \quad -1 < y < 1.$$

    Are X and Y independent?

81. Find $\mathrm{Var}[X]$, $\mathrm{Var}[Y]$, and $\mathrm{Cov}(X,Y)$ if X and Y follow

$$f(x,y) = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}, \quad -\infty < x < \infty, \quad -\infty < y < \infty.$$

82. Consider a sample of $n = 10$ observations displayed in ascending order.

$$15, 16, 18, 18, 20, 20, 21, 22, 23, 75.$$

    (a) Compute the sample mean and sample variance.

    (b) Find the 5-point summary of the data. Is the distribution skewed?

    (c) Are there any likely outliers in the sample? If so, indicate their values.

    (d) Build and display the sample's boxplot chart.

    (e) Build and display a sample histogram.

83. The daily number of accidents in Sydney over a 40-day period are provided below:

6, 3, 2, 24, 12, 3, 7, 14, 21, 9, 14, 22, 15, 2, 17, 10
7, 7, 31, 7, 18, 6, 8, 2, 3, 2, 17, 7, 7, 21, 13, 23, 1, 11
3, 9, 4, 9, 9, 25

(a) Compute the sample mean and sample variance.
(b) Find the 5-point summary of the data. Is the distribution skewed?
(c) Are there any likely outliers in the sample? If so, indicate their values.
(d) Build and display the sample's boxplot chart.
(e) Build and display a sample histogram.

84. Repeat the previous question when the "31" is replaced by a "130".

85. The grades in a class are shown below.

80, 73, 83, 60, 49, 96, 87, 87, 60, 53, 66, 83, 32, 80, 66
90, 72, 55, 76, 46, 48, 69, 45, 48, 77, 52, 59, 97, 76, 89
73, 73, 48, 59, 55, 76, 87, 55, 80, 90, 83, 66, 80, 97, 80
55, 94, 73, 49, 32, 76, 57, 42, 94, 80, 90, 90, 62, 85, 87
97, 50, 73, 77, 66, 35, 66, 76, 90, 73, 80, 70, 73, 94, 59
52, 81, 90, 55, 73, 76, 90, 46, 66, 76, 69, 76, 80, 42, 66
83, 80, 46, 55, 80, 76, 94, 69, 57, 55, 66, 46, 87, 83, 49
82, 93, 47, 59, 68, 65, 66, 69, 76, 38, 99, 61, 46, 73, 90,
66, 100, 83, 48, 97, 69, 62, 80, 66, 55, 28, 83, 59, 48, 61
87, 72, 46, 94, 48, 59, 69, 97, 83, 80, 66, 76, 25, 55, 69
76, 38, 21, 87, 52, 90, 62, 73, 73, 89, 25, 94, 27, 66, 66
76, 90, 83, 52, 52, 83, 66, 48, 62, 80, 35, 59, 72, 97, 69
62, 90, 48, 83, 55, 58, 66, 100, 82, 78, 62, 73, 55, 84, 83
66, 49, 76, 73, 54, 55, 87, 50, 73, 54, 52, 62, 36, 87, 80, 80

(a) Compute the sample mean and sample variance.
(b) Find the 5-point summary of the data. Is the distribution skewed?
(c) Are there any likely outliers in the sample? If so, indicate their values.
(d) Build and display the sample's boxplot chart.
(e) Build and display a sample histogram.
(f) Based on your analysis, how well did the class do?

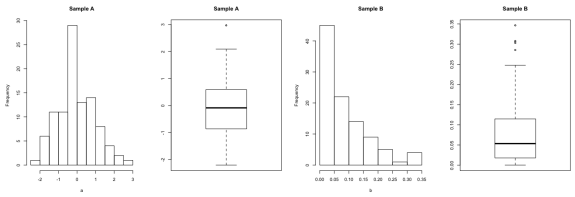
86. Consider the following dataset:

```
2.6   3.7   0.8   9.6   5.8 -0.8   0.7 0.6
4.8   1.2   3.3   5.0   3.7   0.1 -3.1 0.3
```

The median and the interquartile range of the sample are, respectively:

a) 2.4; 3.3 b) 1.9; 3.8 c) 1.9; 1.8 d) 2.9; 12.2
e) NOTP

87. The following charts show a histogram and a boxplot for two samples, A and B. Based on these charts, we may conclude that



a) only A arises from a normal population
b) only B arises from a normal population
c) both A and B arise from a normal population

88. Consider the following dataset:

```
12  14  6  10  1  20  4  8
```

The median and the first quartile of the dataset are, respectively:

a) 9; 5 b) 5.5; 6 c) 10; 5 d) 5; 10 e) NOTP

89. A manufacturer of fluoride toothpaste regularly measures the concentration of of fluoride in the toothpaste to make sure that it is within the specifications of $0.85-1.10$ mg/g.

Table 6.1-3 Concentrations of fluoride in mg/g in toothpaste

0.98	0.92	0.89	0.90	0.94	0.99	0.86	0.85	1.06	1.01
1.03	0.85	0.95	0.90	1.03	0.87	1.02	0.88	0.92	0.88
0.88	0.90	0.98	0.96	0.98	0.93	0.98	0.92	1.00	0.95
0.88	0.90	1.01	0.98	0.85	0.91	0.95	1.01	0.88	0.89
0.99	0.95	0.90	0.88	0.92	0.89	0.90	0.95	0.93	0.96
0.93	0.91	0.92	0.86	0.87	0.91	0.89	0.93	0.93	0.95
0.92	0.88	0.87	0.98	0.98	0.91	0.93	1.00	0.90	0.93
0.89	0.97	0.98	0.91	0.88	0.89	1.00	0.93	0.92	0.97
0.97	0.91	0.85	0.92	0.87	0.86	0.91	0.92	0.95	0.97
0.88	1.05	0.91	0.89	0.92	0.94	0.90	1.00	0.90	0.93

(a) Build a relative frequency histogram of the data (a histogram with area = 1).
(b) Compute the data's mean $\overline{x}$ and its standard deviation s_x.
(c) The mean and the variance can also be approximated as follows. Let u_i be the **class mark** for each of the histogram's classes (the midpoint along the rectangles' widths), n be the total number of observations, and k be the number of classes. Then

$$\overline{u} = \frac{1}{n}\sum_{i=1}^{k} f_i u_i \quad \text{and} \quad s_u^2 = \frac{1}{n-1} f_i (u_i - \overline{u})^2.$$

Compute $\overline{u}$ and s_u. How do they compare with $\overline{x}$ and s_x?
(d) Provide a the 5−point summary of the data, as well as the interquartile range IQR.
(e) Display this information as a boxplot chart.
(f) Compute the **midrange** $\frac{1}{2}(Q_0 + Q_4)$, the **trimean** $\frac{1}{4}(Q_1 + 2Q_2 + Q_3)$, and the **range** $Q_4 - Q_0$ for the fluoride data.

90. Suppose that samples of size $n = 25$ are selected at random from a normal population with mean 100 and standard deviation 10. What is the probability that sample mean falls in the interval

$$(\mu_{\overline{X}} - 1.8\sigma_{\overline{X}}, \mu_{\overline{X}} + 1.0\sigma_{\overline{X}})?$$

91. The amount of time that a customer spends waiting at an airport check-in counter is a random variable with mean $\mu = 8.2$ minutes and standard deviation $\sigma = 1.5$ minutes. Suppose that a random sample of $n = 49$ customers is taken. Compute the approximate probability that the average waiting time for these customers is:

    a) Less than 10 min.

    b) Between 5 and 10 min.

    c) Less than 6 min.

92. A random sample of size $n_1 = 16$ is selected from a normal population with a mean of 75 and standard deviation of 8. A second random sample of size $n_2 = 9$ is taken independently from another normal population with mean 70 and standard deviation of 12. Let $\overline{X}_1$ and $\overline{X}_2$ be the two sample means. Find

    a) The probability that $\overline{X}_1 - \overline{X}_2$ exceeds 4.

    b) The probability that $3.5 < \overline{X}_1 - \overline{X}_2 < 5.5$.

93. Using R, illustrate the central limit theorem by generating $M = 300$ samples of size $n = 30$ from:

    - a normal random variable with mean 10 and variance 0.75;
    - a binomial random variable with 3 trials and probability of success 0.3.

    Repeat the same procedure for samples of size $n = 200$. What do you observe?

94. Suppose that the weight in pounds of a North American adult can be represented by a normal random variable with mean 150 lbs and variance 900 lbs^2. An elevator containing a sign "Maximum 12 people" can safely carry 2000 lbs. The probability that 12 North American adults will not overload the elevator is closest to

    a) 0.97 b) 0.45 c) 0.03 d) 0.00 e) 1.3 f) NOTP

95. Let $X_1, \cdots, X_{50}$ be an independent random sample from a Poisson distribution with mean 1. Set $Y = X_1 + \cdots + X_{50}$. The approximate probability $P(48 \leq Y \leq 52)$ is closest to:

    a) 0.64 b) 0.45 c) 0.22 d) 1.00 e) 0.50 f) NOTP

    A new type of electronic flash for cameras will last an average of 5000 hours with a standard deviation of 500 hours. A quality control engineer intends to select a random sample of 100 of these flashes and use them until they fail. What is the probability that the mean life time of the sample of 100 flashes will be less than 4928 hours?

    a) 0.07 b) 0.93 c) 0.00 d) 0.45 e) NOTP

96. Assume that random variables $\{X_1, \ldots, X_8\}$ follow a normal distribution with mean 2 and variance 24. Independently, assume that random variables $\{Y_1, \ldots, X_{16}\}$ follow a normal distribution with mean 1 and variance 16. Let $\overline{X}$ and $\overline{Y}$ be the corresponding sample means. Then $P(\overline{X} + \overline{Y} > 4)$ is:

    a) 0.77 b) 0.31 c) 0.69 d) 0.99 e) NOTP

97. The compressive strength of concrete is normally distributed with mean $\mu = 2500$ and standard deviation $\sigma = 50$. A random sample of size 5 is taken. What is the standard error of the sample mean?

98. Suppose that $X_1 \sim \mathcal{N}(3, 4)$ and $X_2 \sim \mathcal{N}(3, 45)$. Given that X_1 and X_2 are independent random variables, what is a good approximation to $P(X_1 + X_2 > 9.5)$?

    a) 0.31 b) 0.69 c) 0.53 d) 0.43 e) NOTP

99. A new cure has been developed for a certain type of cement that should change its mean compressive strength. It is known that the standard deviation of the compressive strength is 130 kg/cm^2 and that we may assume that it follows a normal distribution. 9 chunks of cement have been tested and the observed sample mean is $\overline{X} = 4970$. Find the 95% confidence interval for the mean of the compressive strength.

    a) $[4858.37, 5081.63]$ b) $[4885.07, 5054.93]$

    c) $[4858.37, 5054.93]$ d) $[4944.52, 4995.48]$

    e) NOTP

100. Consider the same set-up as in the previous question, but now 100 chunks of cement have been tested and the observed sample mean is $\overline{X} = 4970$. Find the 95% confidence interval for the mean of the compressive strength.

    a) $[4858.37, 5081.63]$ b) $[4885.07, 5054.93]$

    c) $[4858.37, 5054.93]$ d) $[4944.52, 4995.48]$

    e) NOTP

101. Consider the same set-up as in two questions ago, but now we do not know the standard deviation of the normal distribution. 9 chunks of cement have been tested, and the measurements are

    $$5001, 4945, 5008, 5018, 4991, 4990, 4968, 5020, 5003.$$

    Find the 95% confidence interval for the mean of the compressive strength.

    a) $[4858.37, 5081.63]$ b) $[4885.07, 5054.93]$

    c) $[4858.37, 5054.93]$ d) $[4944.52, 4995.48]$

    e) NOTP

102. A steel bar is measured with a device which a known precision of $\sigma = 0.5$mm. Suppose we want to estimate the mean measurement with an error of at most 0.2mm at a level of significance $\alpha = 0.05$. What sample size is required? Assume normality.

    a) 25 b) 24 c) 6 d) 7 e) NOTP

103. In a random sample of 1000 houses in the city, it is found that 228 are heated by oil. Find a 99% C.I. for the proportion of homes in the city that are heated by oil.

a) $[0.202, 0.254]$ b) $[0.197, 0.259]$ c) $[0.194, 0.262]$

d) $[0.185, 0.247]$ e) NOTP

104. Past experience indicates that the breaking strength of yarn used in manufacturing drapery material is normally distributed and that $\sigma = 2$ psi. A random sample of 15 specimens is tested and the average breaking strength is found to be $\overline{x} = 97.5$ psi.

a) Find a 95% confidence interval on the true mean breaking strength.

b) Find a 99% confidence interval on the true mean breaking strength.

105. The diameter holes for a cable harness follow a normal distribution with $\sigma = 0.01$ inch. For a sample of size 10, the average diameter is 1.5045 inches.

a) Find a 99% confidence interval on the mean hole diameter.

b) Repeat this for $n = 100$.

106. A journal article describes the effect of delamination on the natural frequency of beams made from composite laminates. The observations are as follows:

$$230.66, 233.05, 232.58, 229.48, 232.58, 235.22.$$

Assuming that the population is normal, find a 95% confidence interval on the mean natural frequency.

107. A textile fibre manufacturer is investigating a new drapery yarn, which the company claims has a mean thread elongation of $\mu = 12$ kilograms with standard deviation of $\sigma = 0.5$ kilograms.

a) What should be the sample size so that with probability 0.95 we will estimate the mean thread elongation with error at most 0.15 kg?

b) What should be the sample size so that with probability 0.95 we will estimate the mean thread elongation with error at most 0.05 kg?

108. An article in *Computers and Electrical Engineering* considered the speed-up of cellular neural networks (CNN) for a parallel general-purpose computing architecture. Various speed-ups are observed:

```
3.77  3.35  4.21  4.03  4.03  4.63
4.63  4.13  4.39  4.84  4.26  4.60
```

Assume that the population is normally distributed. The 99% C.I. for the mean speed-up is:

a) $[4.155, 4.323]$ b) $[3.863, 4.615]$ c) $[4.040, 4.438]$

d) $[3.77, 4.60]$ e) NOTP

109. An engineer measures the weight of $n = 25$ pieces of steel, which follows a normal distribution with variance 16. The average observed weight for the sample is $\overline{x} = 6$. What is the two-sided 95% C.I. for the mean μ?

110. The brightness of television picture tube can be evaluated by measuring the amount of current required to achieve a particular brightness level. An engineer thinks that one has to use 300 microamps of current to achieve the required brightness level. A sample of size $n = 20$ has been taken to verify the engineer's hypotheses.

a) Formulate the null and the alternative hypotheses (use a two-sided test alternative).

b) For the sample of size $n = 20$ we obtain $\overline{x} = 319.2$ and $s = 18.6$. Test the hypotheses from part a) with $\alpha = 5\%$ by computing a critical region. Calculate the p-value.

c) Use the data from part b) to construct a 95% confidence interval for the mean required current.

111. We say that a particular production process is **stable** if it produces at most 2% defective items. Let p be the true proportion of defective items.

a) We sample $n = 200$ items at random and consider hypotheses testing about p. Formulate null and alternative hypotheses.

b) What is your conclusion of the above test, if one observes 3 defective items out of 200? Note: you have to choose an appropriate confidence level α.

112. Ten engineers' knowledge of basic statistical concepts was measured on a scale of $0-100$, before and after a short course in statistical quality control. The results are:

Engineer	1	2	3	4	5
Before X_{1i}	43	82	77	39	51
After X_{2i}	51	84	74	48	53
Engineer	6	7	8	9	10
Before X_{1i}	66	55	61	79	43
After X_{2i}	61	59	75	82	53

Let μ_1 and μ_2 be the mean mean score before and after the course. Perform the test $H_0 : \mu_1 = \mu_2$ against $H_A : \mu_1 < \mu_2$. Use $\alpha = 0.05$.

113. It is claimed that 15% of a certain population is left-handed, but a researcher doubts this claim. They decide to randomly sample 200 people and use the anticipated small number to provide evidence against the claim of 15%. Suppose 22 of the 200 are left-handed. Compute the p—value associated with the hypothesis (assuming a binomial distribution), and provide an interpretation.

114. A child psychologist believes that nursery school attendance improves children's social perceptiveness (SP). They use 8 pairs of twins, randomly choosing one to attend nursery school and the other to stay at home, and then obtains scores for all 16. In 6 of the 8 pairs, the twin attending nursery school scored better on the SP test. Compute the p—value associated with the hypothesis (assuming a binomial distribution), and provide an interpretation.

115. A certain power supply is stated to provide a constant voltage output of 10kV. Ten measurements are taken and yield the sample mean of 11kV. Formulate a test for this situation. Should it be 1—sided or 2—sided? What value of α should you use? What conclusion does the test and the sample yield?

116. A company is currently using titanium alloy rods it purchases from supplier A. A new supplier (supplier B) approaches the company and offers the same quality (at least according to supplier B's claim) rods at a lower price.

    The company's decision makers are interested in the offer. At the same time, they want to make sure that the safety of their product is not compromised.

    They randomly selects ten rods from each of the lots shipped by suppliers A and B and measures the yield strengths of the selected rods. The observed sample mean and sample standard deviation are 651 MPa and 2 MPa for supplier's A rods, respectively, and the same parameters are 657 MPa and 3 MPa for supplier B's rods.

    Perform the test $H_0 : \mu_A = \mu_B$ against $\mu_A \neq \mu_B$. Use $\alpha = 0.05$. Assume that the variances are equal but unknown.

117. The deflection temperature under load for two different types of plastic pipe is being investigated. Two random samples of 15 pipe specimens are tested, and the deflection temperatures observed are as follows:

    Type 1: 206, 188, 205, 187, 194, 193, 207, 185, 189, 213, 192, 210, 194, 178, 205.

    Type 2: 177, 197, 206, 201, 180, 176, 185, 200, 197, 192, 198, 188, 189, 203, 192.

    Does the data support the claim that the deflection temperature under load for type 1 pipes exceeds that of type 2? Calculate the p-value, using $\alpha = 0.05$, and state your conclusion.

118. It is claimed that the breaking strength of yarn used in manufacturing drapery material is normally distributed with mean 97 and $\sigma = 2$ psi. A random sample of nine specimens is tested and the average breaking strength is found to be $\overline{X} = 98$ psi. Formulate a test for this situation. Should it be 1—sided or 2—sided? What value of α should you use? What conclusion does the test and the sample yield?

119. A civil engineer is analyzing the compressive strength of concrete. It is claimed that its mean is 80 and variance is known to be 2. A random sample of size 60 yields the sample mean 59. Formulate a test for this situation. Should it be 1—sided or 2—sided? What value of α should you use? What conclusion does the test and the sample yield?

120. The sugar content of the syrup in canned peaches is claimed to be normally distributed with mean 10 and variance 2. A random sample of $n = 10$ cans yields a sample mean 11. Another random sample of $n = 10$ cans yields a sample mean 9. Formulate a test for this situation. Should it be 1—sided or 2—sided? What value of α should you use? What conclusion does the test and the sample yield?

121. The mean water temperature downstream from a power water plant cooling tower discharge pipe should be no more than 100F. Past experience has indicated that that the standard deviation is 2F. The water temperature is measured on nine randomly chosen days, and the average temperature is found to be 98F. Formulate a test for this situation. Should it be 1—sided or 2—sided? What value of α should you use? What conclusion does the test and the sample yield?

122. We are interested in the mean burning rate of a solid propellant used to power aircrew escape systems. We want to determine whether or not the mean burning rate is 50 cm/second. A sample of 10 specimens is tested and we observe $\overline{X} = 48.5$. Assume normality with $\sigma = 2.5$.

123. Ten individuals have participated in a diet modification program to stimulate weight loss. Their weight both before and after participation in the program is shown below:

Before	195, 213, 247, 201, 187, 210, 215, 246, 294, 310
After	187, 195, 221, 190, 175, 197, 199, 221, 278, 285

    Is there evidence to support the claim that this particular diet-modification program is effective in producing mean weight reduction? Use $\alpha = 0.05$. Compute the associated p—value.

124. We want to test the hypothesis that the average content of containers of a particular lubricant equals 10L against the two-sided alternative. The contents of a random sample of 10 containers are

10.2	9.7	10.1	10.3	10.1
9.8	9.9	10.4	10.3	9.5

    Find the p—value of this two-sided test. Assume that the distribution of contents is normal. Note that if x_i represent the measurements, $\sum_{i=1}^{10} x_i^2 = 1006.79$.

    a) $0.05 < p < 0.10$ b) $0.10 < p < 0.20$ c) $0.25 < p < 0.40$

    d) $0.50 < p < 0.80$ e) NOTP

125. An engineer measures the weight of $n = 25$ pieces of steel, which follows a normal distribution with variance 16. The average weight for the sample is $\overline{X} = 6$. They want to test for $H_0 : \mu = 5$ against $H_1 : \mu > 5$. What is the p—value for the test?

    a) 0.05 b) 0.11 c) 0.89 d) 1.00 e) NOTP

126. The thickness of a plastic film (in mm) on a substrate material is thought to be influenced by the temperature at which the coating is applied. A completely randomized experiment is carried out. 11 substrates are coated at 125F, resulting in a sample mean coating thickness of $\overline{x}_1 = 103.5$ and a sample standard deviation of $s_1 = 10.2$. Another 11 substrates are coated at 150F, for which $\overline{x}_2 = 99.7$ and $s_2 = 11.7$ are observed. We want to test equality of means against the two-sided alternative. Assume that population variances are unknown but equal. The value of the appropriate test statistics and the decision are (for $\alpha = 0.05$):

    a) 0.81; Reject H_0. b) 0.81; Do not reject H_0.

    c) 1.81; Reject H_0. d) 1.81; Do not reject H_0.

    e) NOTP

127. The following output was produced with `t.test` command in `R`.

```
One Sample t-test
data:  x
t = 2.0128, df = 99, p-value = 0.02342
alternative hypothesis: true mean is greater th
```

Based on this output, which statement is correct?

a) If the type I error is 0.05, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu > 0$;

b) If the type I error is 0.05, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu \neq 0$;

c) If the type I error is 0.01, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu > 0$;

d) If the type I error is 0.01, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu < 0$;

e) Type I error is 0.02342.

128. A pharmaceutical company claims that a drug decreases a blood pressure. A physician doubts this claim. They test 10 patients and records results before and after the drug treatment:

```
> Before=c(140,135,122,150,126,
    138,141,155,128,130)
> After=c(135,136,120,148,122,
    136,140,153,120,128)
```

At the R command prompt, they type:

```
> test.t(Before,After,alternative=
    "greater")
    data:  Before and After
    t = 0.5499, p-value = 0.2946
    alternative hypothesis: true
        difference in means is
        greater than 0
    sample estimates: mean of x mean of y
        136.5     133.8
```

Their assistant claims that the command should instead be:

```
> test.t(Before,After,paired=TRUE,
    alternative="greater")

    data: Before and After t = 3.4825,
        df = 9, p-value = 0.003456
    alternative hypothesis: true
        difference in means is
        greater than 0
    sample estimates: mean of the
        differences
        2.7
```

Which answer is best?

a) The assistant uses the correct command. There is not enough evidence to justify that the new drug decreases blood pressure;

b) The assistant uses the correct command. There is enough evidence to justify that the new drug decreases blood pressure for any reasonable choice of α;

c) The physician uses the correct command. There is not enough evidence to justify that the new drug decreases blood pressure;

d) The physician uses the correct command. There is enough evidence to justify that the new drug decreases blood pressure for any reasonable choice of α;

e) Nobody is correct, t—tests should not be used here.

129. A company claims that the mean deflection of a piece of steel which is 10ft long is equal to 0.012ft. A buyer suspects that it is bigger than 0.012ft. The following data x_i has been collected:

$$0.0132, 0.0138, 0.0108, 0.0126, 0.0136,$$
$$0.0112, 0.0124, 0.0116, 0.0127, 0.0131.$$

Assuming normality and that $\sum_{i=1}^{10} x_i^2 = 0.0016$, what are the p—value for the appropriate one-sided test and the corresponding decision?

a) $p \in (0.05, 0.1)$ and reject H_0 at $\alpha = 0.05$.

b) $p \in (0.05, 0.1)$ and do not reject H_0 at $\alpha = 0.05$.

c) $p \in (0.1, 0.25)$ and reject H_0 at $\alpha = 0.05$.

d) $p \in (0.1, 0.25)$ and do not reject H_0 at $\alpha = 0.05$.

130. In an effort to compare the durability of two different types of sandpaper, 10 pieces of type A sandpaper were subjected to treatment by a machine which measures abrasive wear; 11 pieces of type B sandpaper were subjected to the same treatment. We have the following observations:

```
xA 27 26 24 29 30 26 27 23 28 27
xB 24 23 22 27 24 21 24 25 24 23 20
```

Note that $\sum x_{A,i} = 267$, $\sum x_{B,i} = 257$, $\sum x_{A,i}^2 = 7169$, $\sum x_{B,i}^2 = 6041$. Assuming normality and equality of variances in abrasive wear for A and B, we want to test for equality of mean abrasive wear for A and B. The appropriate p—value is

a) $p < 0.01$ b) $p > 0.2$ c) $p \in (0.01, 0.05)$
d) $p \in (0.1, 0.2)$ e) $p \in (0.05, 0.1)$ f) NOTP

131. The following output was produced with `t.test` command in R.

```
One Sample t-test
data:  x
t = 32.9198, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not
    equal to 0
```

Based on this output, which statement is correct?

a) If the type I error is 0.05, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu > 0$;

b) If the type I error is 0.05, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu \neq 0$;

c) If the type I error is 0.01, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu > 0$;

d) If the type I error is 0.01, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu < 0$;

e) NOTP

132. Consider a sample $\{X_1, \ldots, X_{10}\}$ from a normal population $X_i \sim \mathcal{N}(4, 9)$. Denote by $\overline{X}$ and S^2 the sample mean and the sample variance, respectively. Find c such that

$$P\left(\frac{\overline{X} - 4}{S/\sqrt{10}} \leq c\right) = 0.99$$

a) 1.833 b) 2.326 c) 1.645 d) 2.821 e) NOTP

133. A medical team wants to test whether a particular drug decreases diastolic blood pressure. Nine people have been tested. The team measured blood pressure before (X) and after (Y) applying the drug. The corresponding means were $\overline{X} = 91$, $\overline{Y} = 87$. The sample variance of the differences was $S_D^2 = 25$. The $p-$value for the appropriate one-sided test is between:

a) 0 and 0.025 b) 0.025 and 0.05 c) 0.05 and 0.1

d) 0.1 and 0.25 e) 0.25 and 1 f) NOTP

134. A researcher studies a difference between two programming languages. Twelve experts familiar with both languages were asked to write a code for a particular function using both languages and the time for writing those codes was registered. The observations are as follows.

```
Expert 01 02 03 04 05 06 07 08 09 10 11 12
Lang 1 17 16 21 14 18 24 16 14 21 23 13 18
Lang 2 18 14 19 11 23 21 10 13 19 24 15 29
```

Construct a 95% C.I. for the mean difference between the first and the second language. Do we have any evidence that one of the languages is preferable to the other (i.e. the average time to write a function is shorter)?

a) $[-1.217, 2.550]$, indication that language 2 is better

b) $[-1.217, 2.550]$, no evidence that any of them is better

c) $[-1.217, 2.550]$, indication that language 1 is better

d) $[-2.86, 4.19]$, no evidence that any of them is better

135. Consider a proportion of recaptured moths in the light-coloured (p_1) and the dark-coloured (p_2) populations.

Among the $n_1 = 137$ light-coloured moths, $y_1 = 18$ were recaptured; among the $n_2 = 493$ dark-coloured moths, $y_2 = 131$ were recaptured. Is there a significant difference between the proportion of recaptured moths in both populations?

136. For a set of 12 pairs of observations on (x_i, y_i) from an experiment, the following summary for x and y is obtained:

$$\sum_{i=1}^{12} x_i = 25, \sum_{i=1}^{12} y_i = 432,$$
$$\sum_{i=1}^{12} x_i^2 = 59, \sum_{i=1}^{12} x_i y_i = 880.5, \sum_{i=1}^{12} y_i^2 = 15648.$$

The estimated value of y at $x = 5$ from the least squares regression line is:

a) 27.78 b) 47.77 c) 41.87 d) 55.97 e) NOTP

137. Assuming that the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ is appropriate for $n = 14$ observations, the estimated regression line is computed to be

$$\hat{y} = 0.66490 + 0.83075x.$$

Given that $S_{yy} = 4.1289$ and $S_{xy} = 4.49094$, compute the estimated standard error for the slope.

a) 0.32 b) 0.08 c) 0.09 d) 0.01 e) NOTP

138. We have a dataset with $n = 25$ pairs of observations (x_i, y_i), and

$$\sum_{i=1}^{n} x_i = 325.000, \sum_{i=1}^{n} y_i = 658.972,$$
$$\sum_{i=1}^{n} x_i^2 = 5525.000, \sum_{i=1}^{n} x_i y_i = 11153.588,$$
$$\sum_{i=1}^{n} y_i^2 = 22631.377.$$

Note that $t_{0.05/2}(23) = 2.069$. The point estimate for the slope of the regression line is

a) 1.99 b) -1.99 c) 0.49 d) 0.59 e) NOTP

139. Use the same data as in the previous question. What is the point estimate for the intercept of the regression line?

a) 1.99 b) -1.99 c) 0.49 d) 0.59 e) NOTP

140. Use the same data as in the previous question. What is the prediction of y for $x = 30$?

a) 60.19 b) 16.67 c) 30 d) 30.54 e) NOTP

141. Use the same data as in the previous question. Note that $t_{0.05/2}(23) = 2.069$. Is the linear regression significant?

142. A company employs 10 part-time drivers for its fleet of trucks. Its manager wants to find a relationship between number of km driven (X) and number of working days (Y) in a typical week. The drivers are hired to drive half-day shifts, so that 3.5 stands for 7 half-day shifts.

The manager wants to use the linear regression model $Y = \beta_0 + \beta_1 x + \epsilon$ on the following data:

	1	2	3	4	5
x	825	215	1070	550	480
y	3.5	1.0	4.0	2.0	1.0
	6	7	8	9	10
x	920	1350	325	670	1215
y	3.0	4.5	1.5	3.0	5.0

Note that $\sum x_i^2 = 7104300$, $\sum y_i^2 = 99.75$, and $\sum x_i y_i = 26370$. What is the fitted regression line?

143. Using the data from the previous question, what value is the correlation coefficient of x and y closest to?

a) 0.44 b) 0.95 c) 0.11 d) 1.12 e) NOTP

144. We want to test significance of regression, i.e. $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. The value of the appropriate statistic and the decision for $\alpha = 0.05$ is:

a) 8.55; do not reject H_0 b) 2.31; reject H_0

c) 8.55; reject H_0 d) 2.31; do not reject H_0

e) NOTP

145. Regression methods were used to analyze the data from a study investigating the relationship between roadway surface temperature in F (x) and pavement defection (y). Summary quantities were $n = 20$,

$$\sum y_i = 12.75,\ \sum y_i^2 = 8.86,$$
$$\sum x_i = 1478\ \sum x_i^2 = 143,215.8\ \sum x_i y_i = 1083.67.$$

a) Calculate the least squares estimates of the slope and intercept. Estimate σ^2.

b) Use the equation of the fitted line to predict what pavement deflection would be observed when the surface temperature is 90F.

c) Give a point estimate of the mean pavement deflection when the surface is 85F.

d) What change in mean pavement deflection would be expected for a 1F change in surface temperature?

146. Consider the data from the previous question.

a) Test for significance of regression using $\alpha = 0.05$. Find the p-value for this test. What conclusion can you draw?

b) Estimate the standard errors of the slope and intercept.

147. Solve this question using `R`.

a) Generate a sample `x` of size $n = 100$ from a normal distribution;

b) Define `y=1+2*x+rnorm(100);`

c) Plot scatter plot;

d) Find the estimators of the regression parameters and add the line to the scatter plot;

f) Compute the correlation coefficient

g) Plot the residuals;

h) Comment on your results.

148. We have a dataset with $n = 10$ pairs of observations (x_i, y_i), and

$$\sum_{i=1}^{n} x_i = 683,\ \sum_{i=1}^{n} y_i = 813,$$
$$\sum_{i=1}^{n} x_i^2 = 47,405,\ \sum_{i=1}^{n} x_i y_i = 56,089,\ \sum_{i=1}^{n} y_i^2 = 66,731.$$

a) What is the line of best fit for this data?

b) What is an approximate 95% confidence interval for the intercept and the slope of the line of best fit?

c) What is an approximate 95% confidence interval for the mean response at $x_0 = 60$? At $x_0 = 90$?

d) What is an approximate 95% prediction interval for the response y_0 at $x_0 = 60$? At $x_0 = 90$?

e) What is the mean squared error estimate for the variance of the residuals?

149. Repeat parts b), c), and d) from the previous questions by using 99% instead of 95%.

References

[1] Y. Cissokho, S. Fadel, R. Millson, R. Pourhasan, and P. Boily. Anomaly Detection and Outlier Analysis. *Data Science Report Series*, 2020.

[2] U. Dudley. *Mathematical Cranks*. Mathematical Association of America, 1992.

[3] E. Ghashim and P. Boily. A Soft Introduction to Bayesian Data Analysis. *Data Science Report Series*, 2020.

[4] E. Gibson. The role of p—values in judging the strength of evidence and realistic replication expectations. *Statistics in Biopharmaceutical Research*, 13(1):6–18, 2021.

[5] A. K. Han. Non-parametric analysis of a generalized regression model. *J. Econometrics*, 35:303–316, 1987.

[6] R. Hogg, J. McKean, and A. Craig. *Introduction to Mathematical Statistics*. Pearson, 6th edition, 2005.

[7] R. Hogg and E. Tanis. *Probability and Statistical Inference*. Pearson/Prentice Hall, 7th edition, 2006.

[8] M. Hollander and D. Wolfe. *Nonparametric Statistical Methods*. Wiley, 2nd edition, 1999.

[9] E. Jaynes. *Probability Theory: the Logic of Science*. Cambridge Press, 2003.

[10] P. John. *Statistical Design and Analysis of Experiments*. SIAM, 1971.

[11] R. Johnson and D. Wichern. *Applied multivariate statistical analysis*. Prentice Hall, Upper Saddle River, NJ, 5th edition, 2002.

[12] R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Pearson, 6th edition, 2007.

[13] M. Kutner, C. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw Hill Irwin, 2004.

[14] Mathematical Association, UK. An Aeroplane's Guide to A Level Maths.

[15] D. Montgomery. *Introduction to Mathematical Statistics*. Wiley, 7th edition, 2009.

[16] A. Reinhart. *Statistics Done Wrong: the Woefully Complete Guide*. No Starch Press, 2015.

[17] S. M. Ross. *Introduction to Probability Models*. Academic Press, San Diego, CA, USA, sixth edition, 1997.

[18] H. Sahai and M. Ageel. *The Analysis of Variance: Fixed, Random and Mixed Models*. Birkhäuser, 2000.

[19] H. Scheffe. *Analysis of variance*. John Wiely and Sons Inc., London, 1959.

[20] D. Sivia and J. Skilling. *Data Analysis: A Bayesian Tutorial (2nd ed.)*. Oxford Science, 2006.

[21] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye. *Probability & statistics for engineers and scientists*. Pearson Education, Upper Saddle River, 8th edition, 2007.

[22] Wikipedia. List of Probability Distributions. 2021.